« Better Data for Better Science »

The challenge of linking health databases

Erwan Drézen
Founder CEO

Cubr

Qu'est-ce que Cubr ?

Une startup rennaise avec une vision spécifique des données de Santé

Des algorithmes ultra véloces traitant de larges volumes de données

Des cas d'usage rendus possibles

Exploration de cohortes, appariement de bdd...

Une expertise sur le SNDS

le Système National des Données de Santé

Les bases de données en Santé

Des données pour la recherche

Ré-utilisation des données de patients

- Pour la recherche (ex: épidémiologie)
- Pour améliorer la politique de Santé Publique

Aspects réglementaires

- Très contrôlé !
- Accord CNIL nécessaire
- Pseudonymisation des données
- Espaces de stockage sécurisés

Les sources de données

Producteurs de données

- Etablissements Hospitaliers
- Médecins, laboratoires, ...
- Assurance Maladie
- Etc, ...

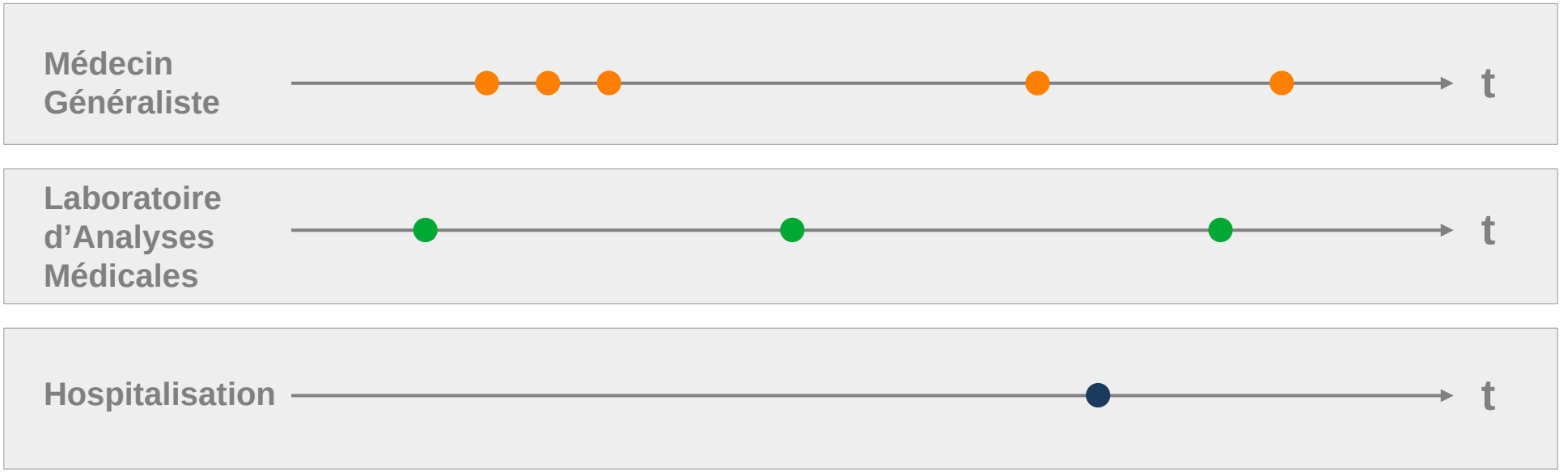
Des sources cloisonnées

- Des espaces protégés (données sensibles !)
- Des structures de données hétérogènes

Des informations disséminées

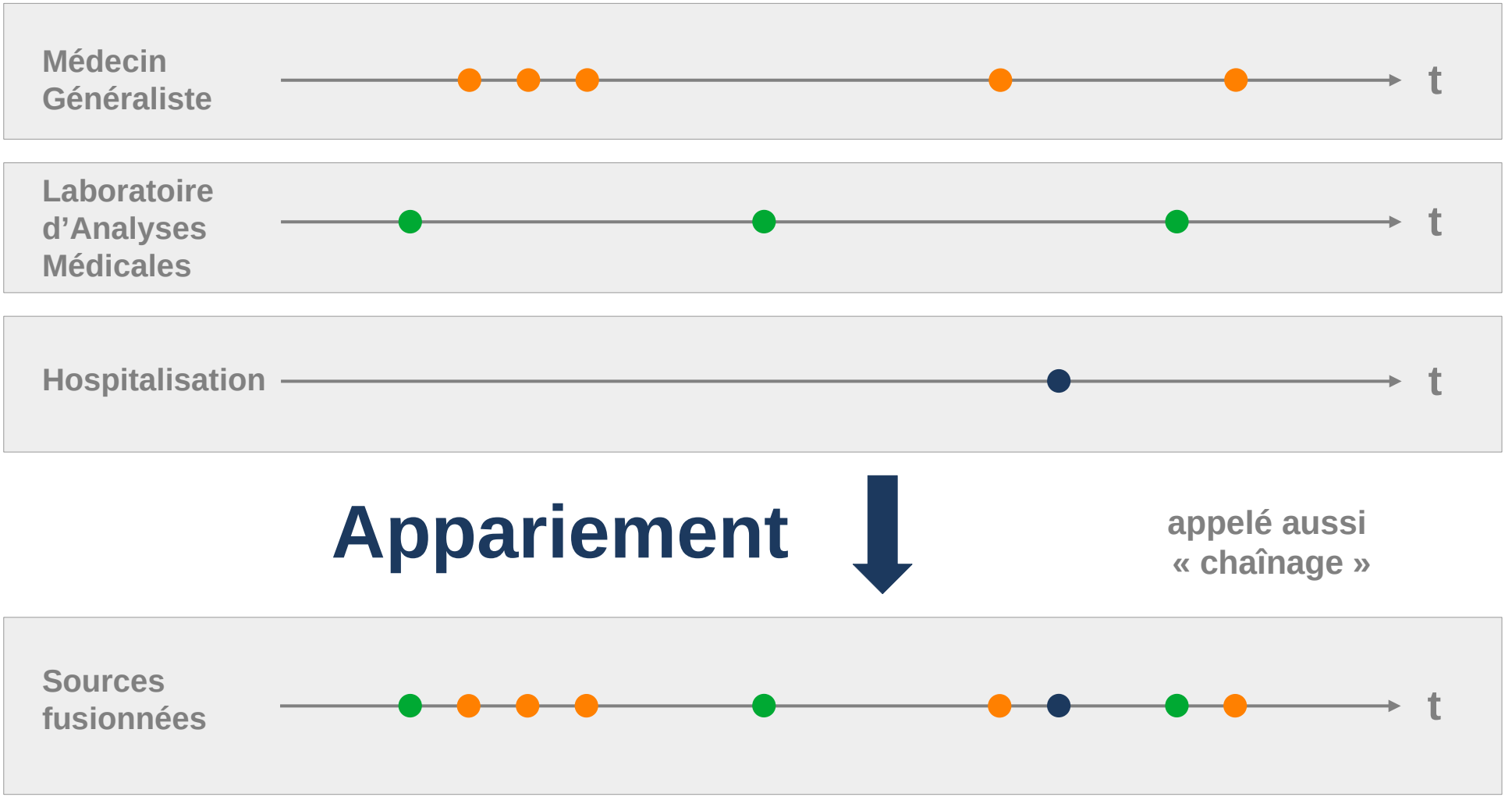
- Parcours de soins morcelés

Des informations disséminées



Les sources de données

Des informations disséminées



Le SNDS, qu'est-ce que c'est ?

Le SNDS, qu'est-ce que c'est ?

Système National des Données de Santé

=

Données de remboursement de l'Assurance Maladie

Quasi totalité de la population française

- Données individuelles
- **Vous êtes très vraisemblablement dedans !**

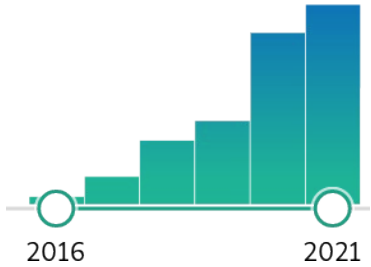
Une base de données volumineuse

- De l'ordre de 200 Térabytes
- Données structurées, des centaines de tables

Le SNDS, qu'est-ce que c'est ?

Ré-utilisation en épidémiologie

- Effectif depuis quelques années
- PubMed : 133 résultats (« SNDS france»)



Avantages

- Quasi exhaustivité de la population
- Historique important (≈2007)
- Possibilité d'études médico-économiques

Inconvénients

- Données de remboursements !
- Aspects réglementaires encore lourds
- Courbe d'apprentissage non négligeable

Quel contenu « exploitable » ?

Données socio-démographiques

- Sexe, dates (naissance/décès), localisations
- Restrictions (données sensibles)

Données « médicales »

- Consultations MG/MS
- Délivrances médicaments
- Actes médicaux, biologies
- Hospitalisations
- Pas de diagnostics (sauf motif hospit. & ALD)
- Pas de données biologiques

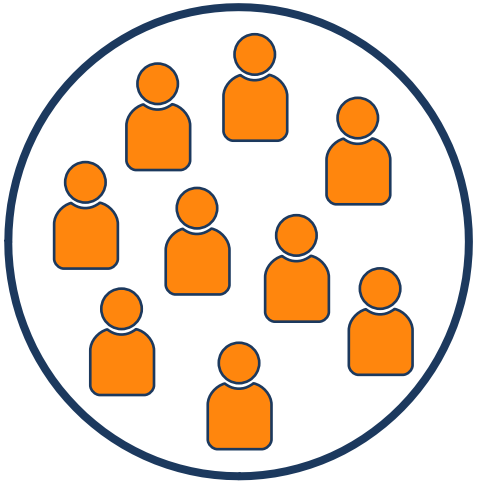
Données économiques

- Tous les montants de remboursements

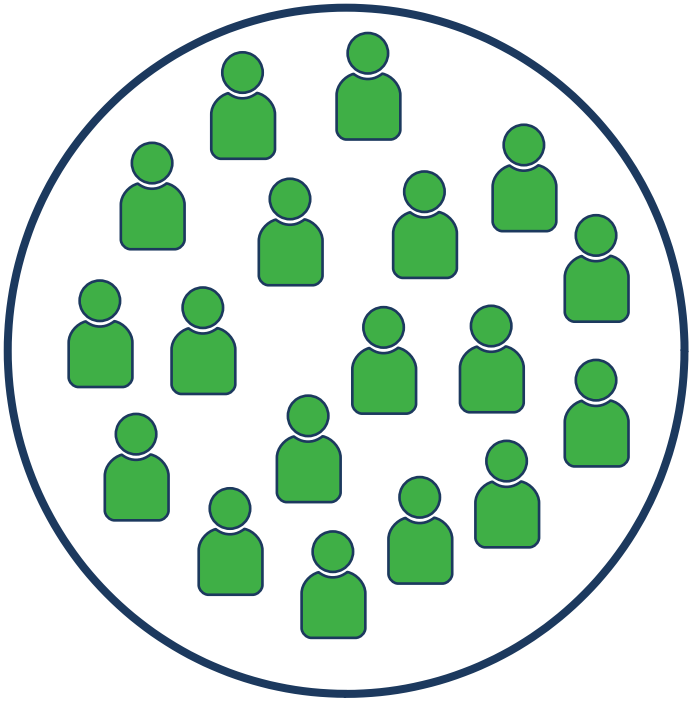
Comment réaliser un appariement avec le SNDS ?

L’appariement & le SNDS

essai clinique



SNDS



Dossier patient
• Données socio-démographiques • Type / origine / taille de tumeur • Diagnostic • IRM (date) • Chirurgie (date) • RCPs • Comorbidités • Etc...

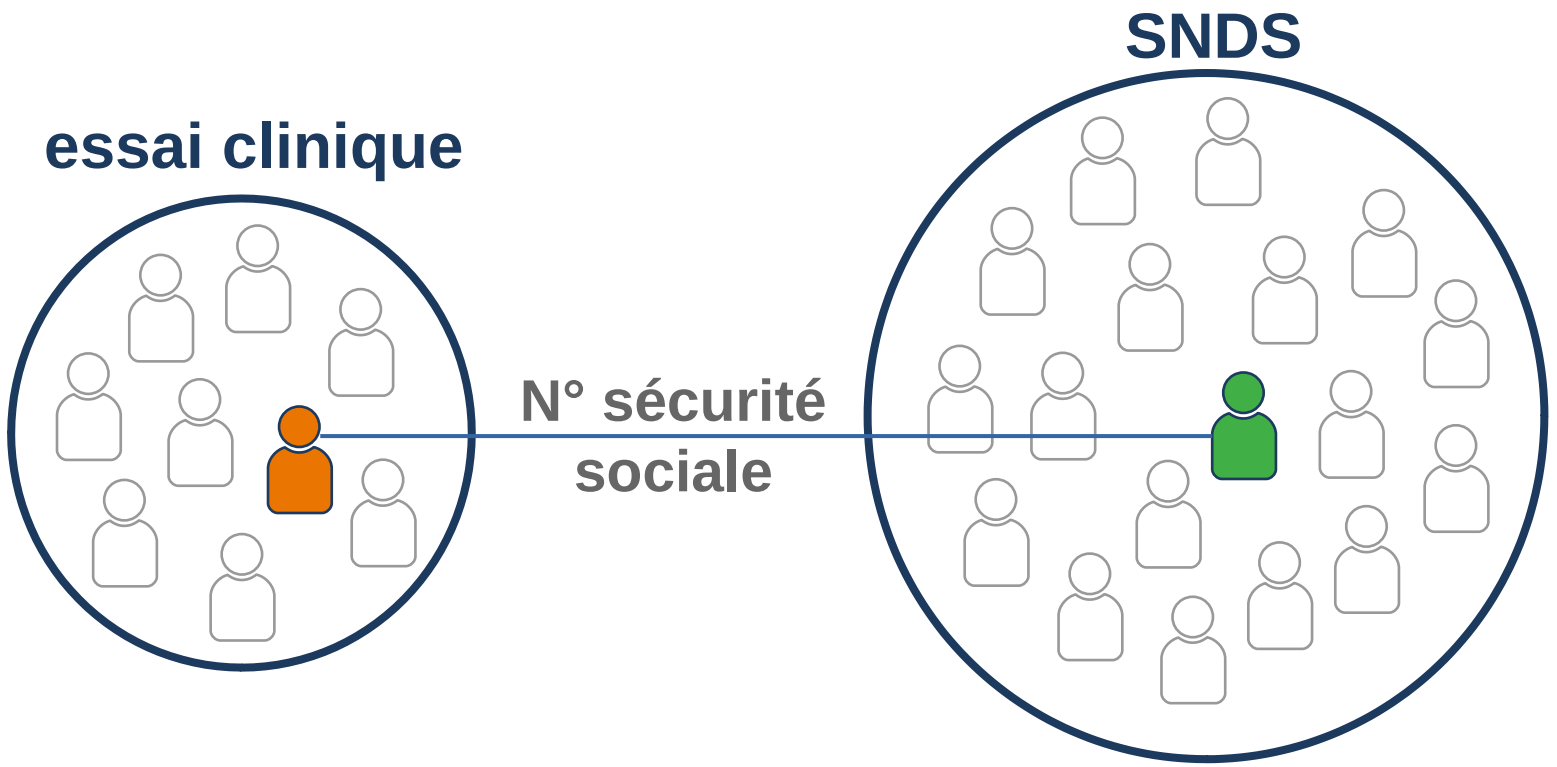
Données SNDS
• Données socio-démographiques • Coûts de remboursements • Consultations MG, MS • Délivrances médicaments • Actes médicaux • Biologies • Hospitalisations PMSI • Etc...

Apparier, comment faire ?

Appariement « direct » : NIR disponible

👍 appariement sûr & simple à réaliser

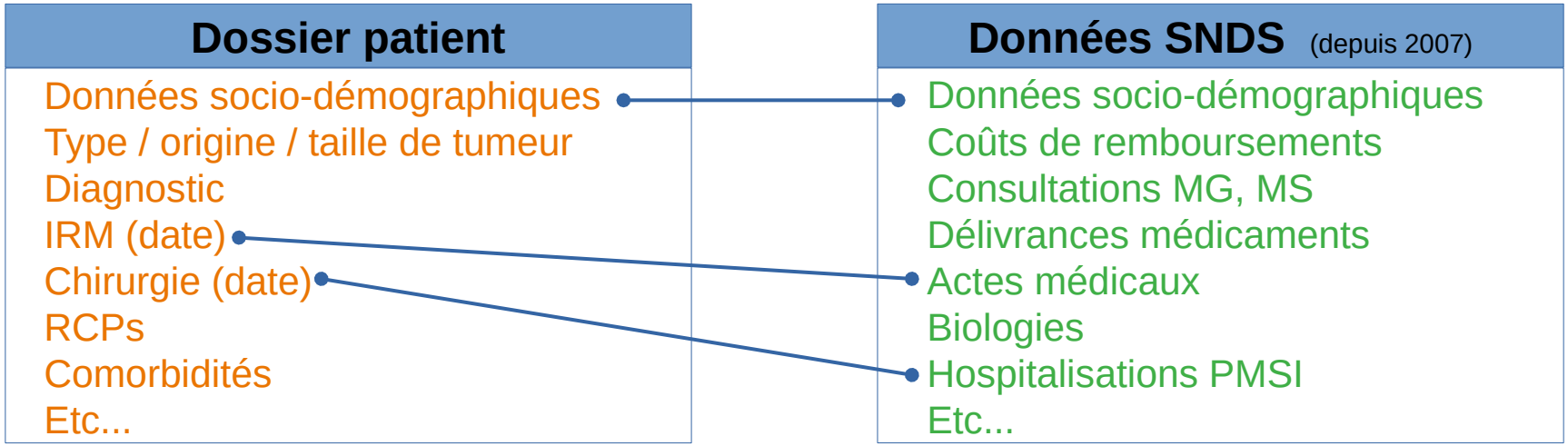
👎 NIR (presque) jamais disponible



Apparier, comment faire ?

Appariement « indirect » : pas de NIR

🤞 tenter de créer un identifiant unique
à partir d'informations communes

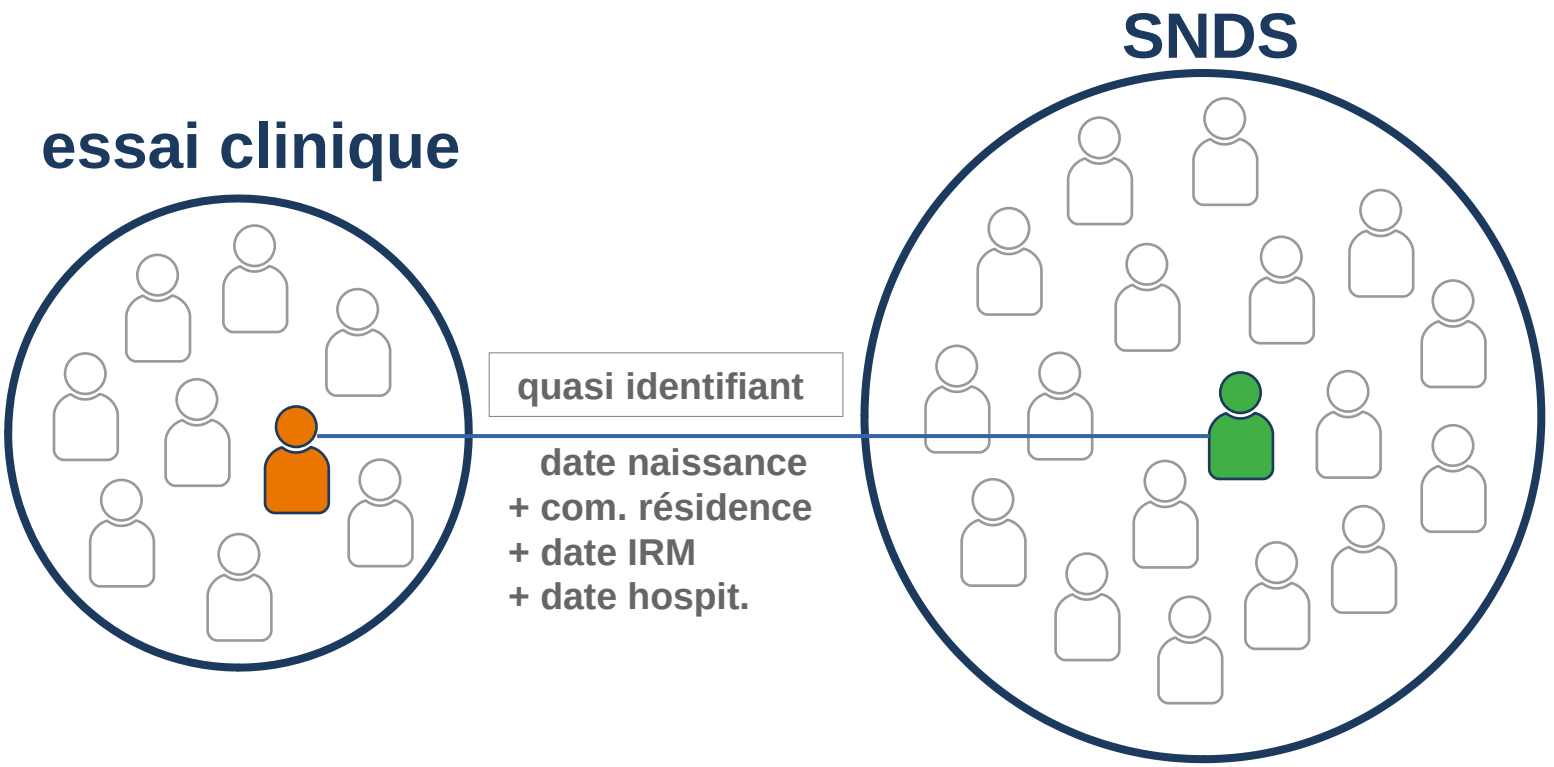


Apparier, comment faire ?

Appariement « indirect » : pas de NIR

🤞 QID : quasi identifiant (qu'on espère unique)

👎 appariement moins sûr & plus complexe



Apparier, comment faire ?

Appariement « indirect » : pas de NIR

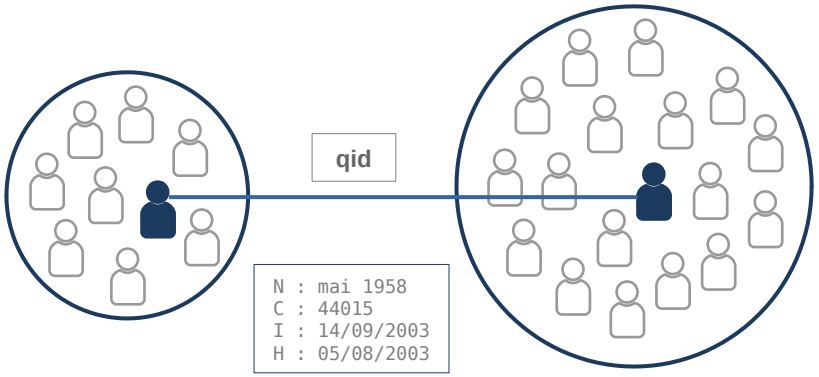
Les ennuis arrivent à grands pas !

Erreurs / incohérences dans les données

Difficultés dans le mapping sémantique

Apparier, comment faire ?

3 cas possibles



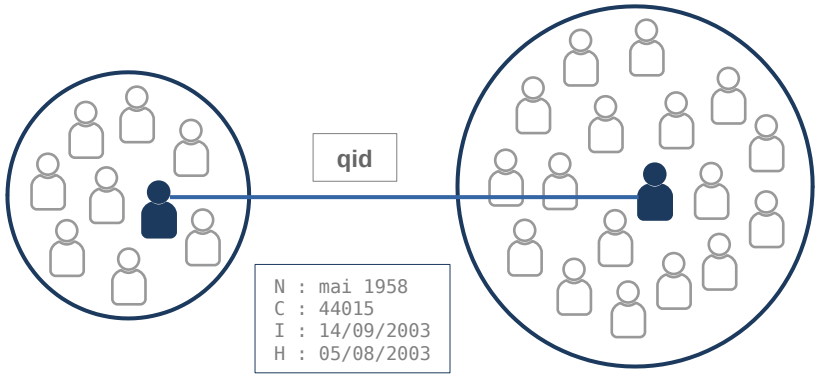
Patient unique trouvé

- Chaque information du qid utilisée (« perfect match »)

🙂 **Patient appariable**

Apparier, comment faire ?

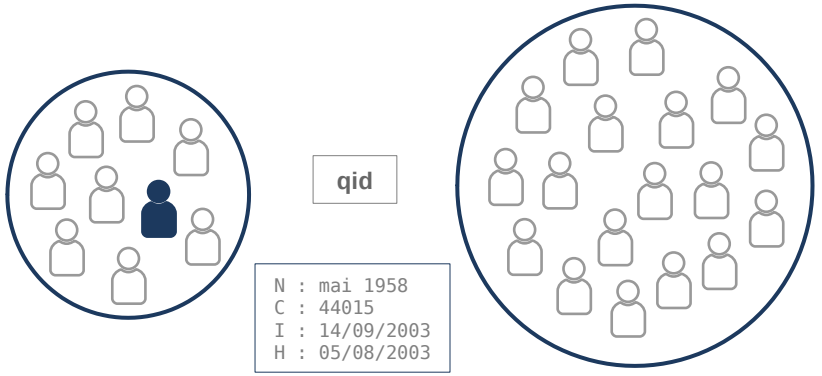
3 cas possibles



Patient unique trouvé

- Chaque information du qid utilisée (« perfect match »)

🙂 **Patient appariable**



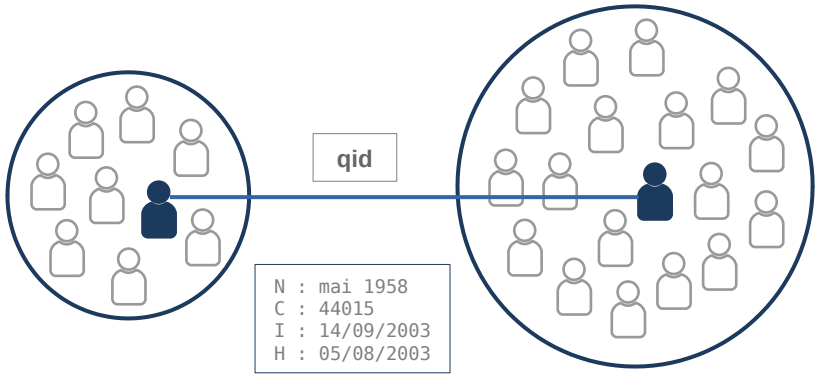
Pas de patient trouvé

- Incohérence dans les données ou
- Incohérence dans le mapping ou
- Patient non présent

☹️ **Patient non appariable**

Apparier, comment faire ?

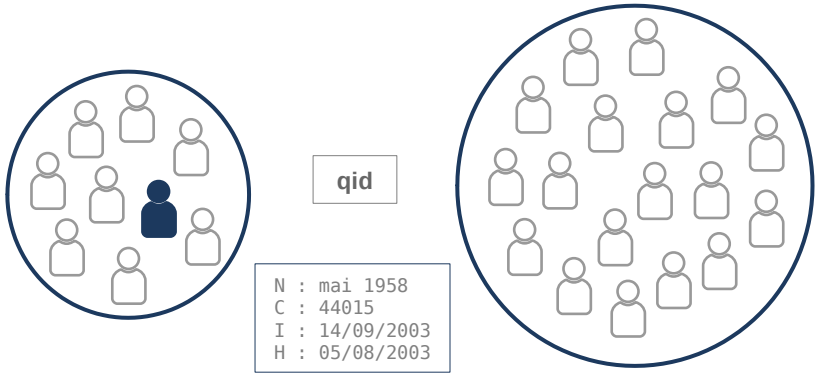
3 cas possibles



Patient unique trouvé

- Chaque information du qid utilisée (« perfect match »)

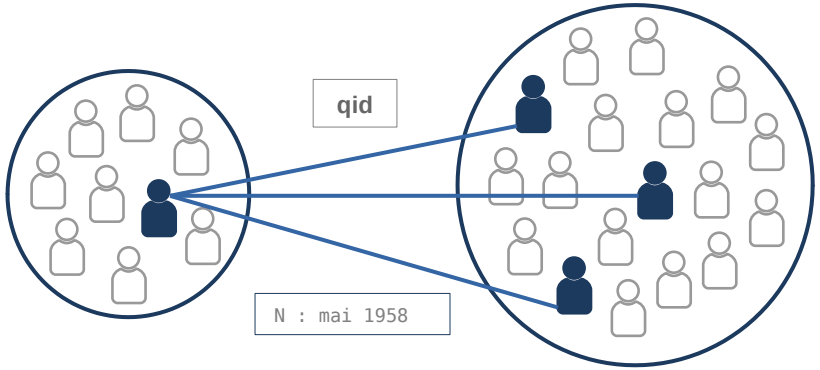
🙂 **Patient appariable**



Pas de patient trouvé

- Incohérence dans les données ou
- Incohérence dans le mapping ou
- Patient non présent

☹️ **Patient non appariable**



Candidats multiples

- qid pas assez discriminant

☹️ **Patient non appariable**

Apparier, comment faire ?

Appariement indirect

Souvent utilisé pour apparier avec le SNDS

- Essai clinique / registre / cohorte avec le SNDS

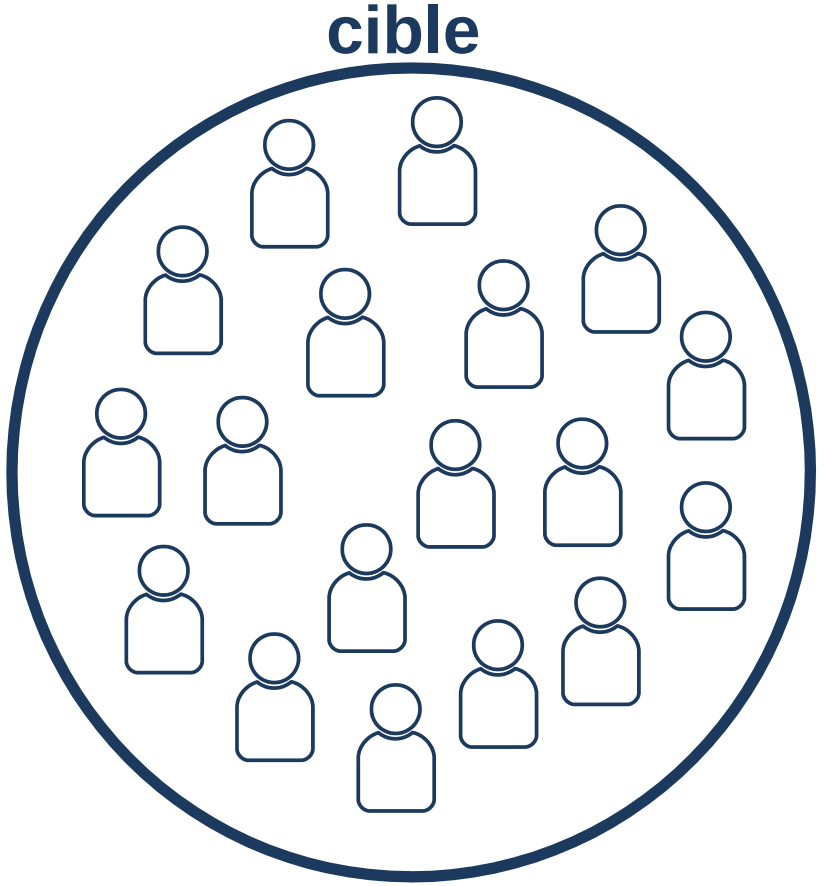
Voué à des résultats mitigés

- Définition d'un seul QID (ou quelques uns)
  - Potentiellement faible % d'appariement

- Confiance dans les résultats ?
  - Peu de contrôle possible *a posteriori*

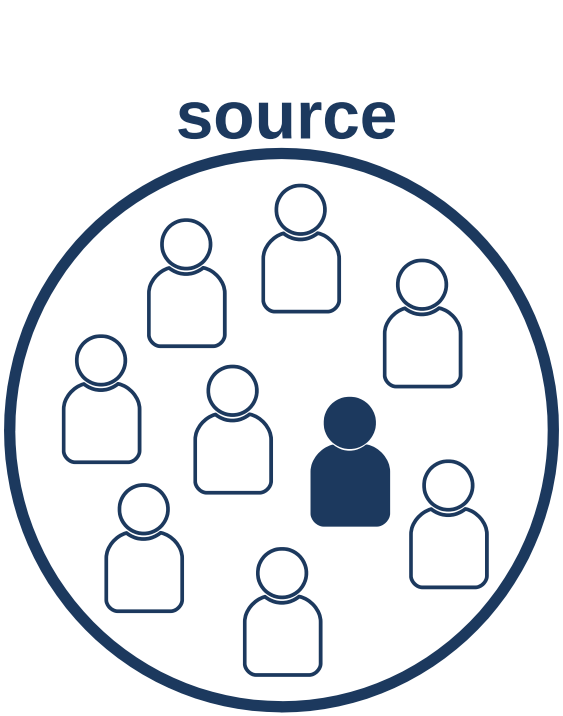
Comment aller plus loin ?

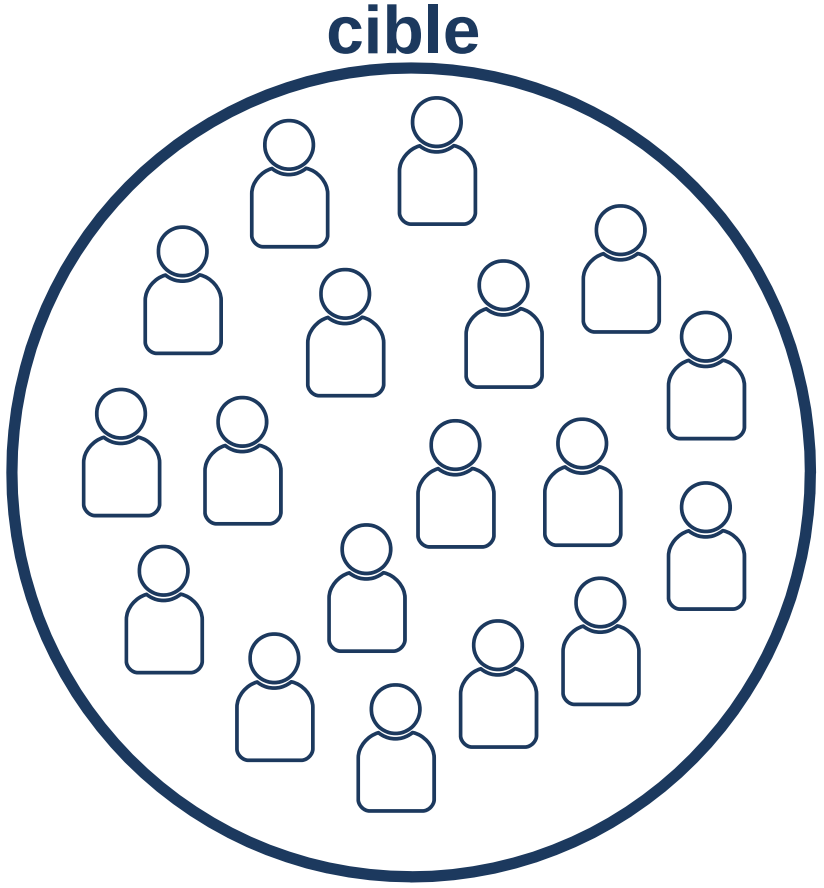
Exemple d'appariement indirect



QID
N : naissance D : décès H : hospitalisation C : code postal

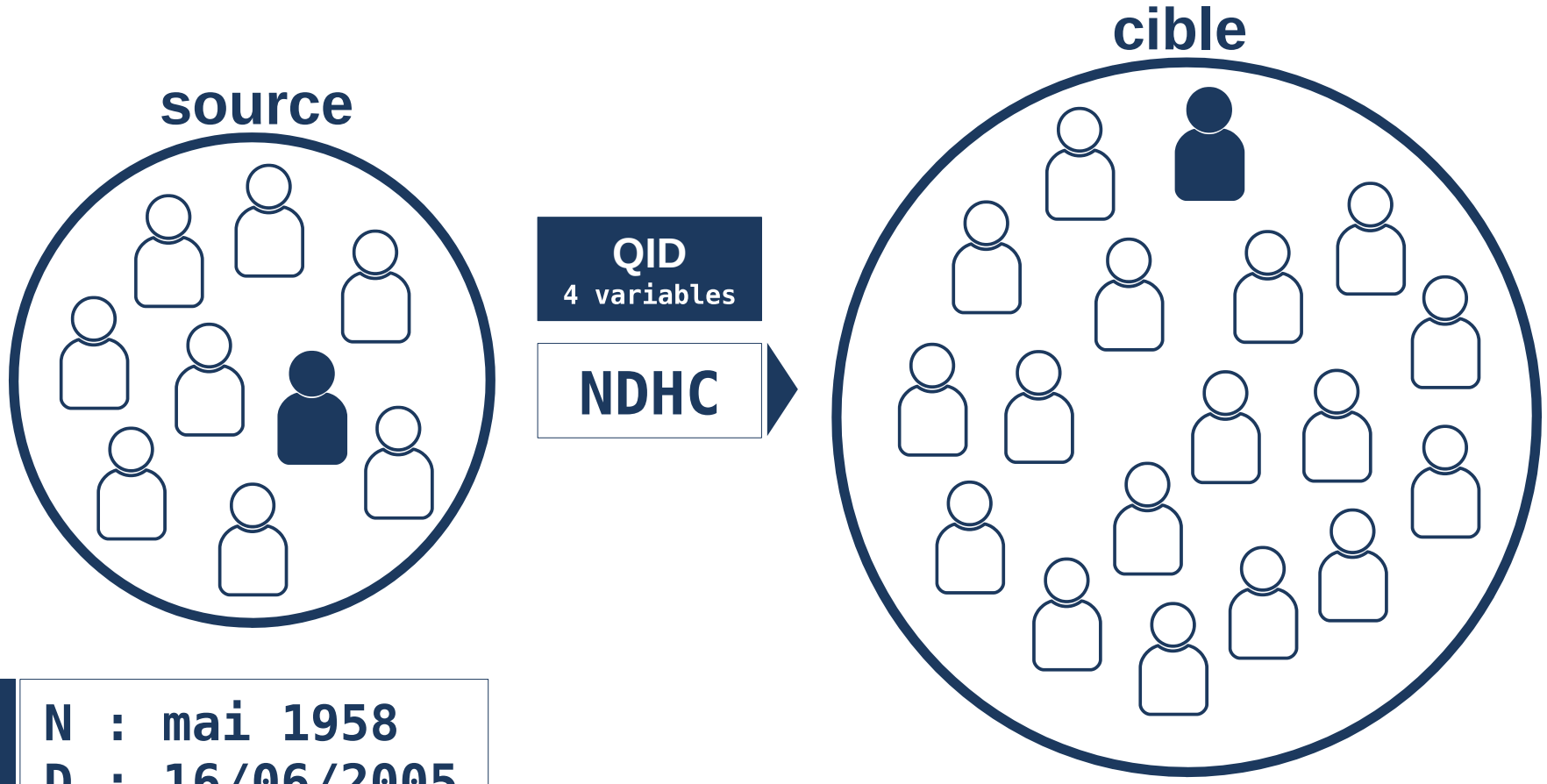
Exemple d'appariement indirect





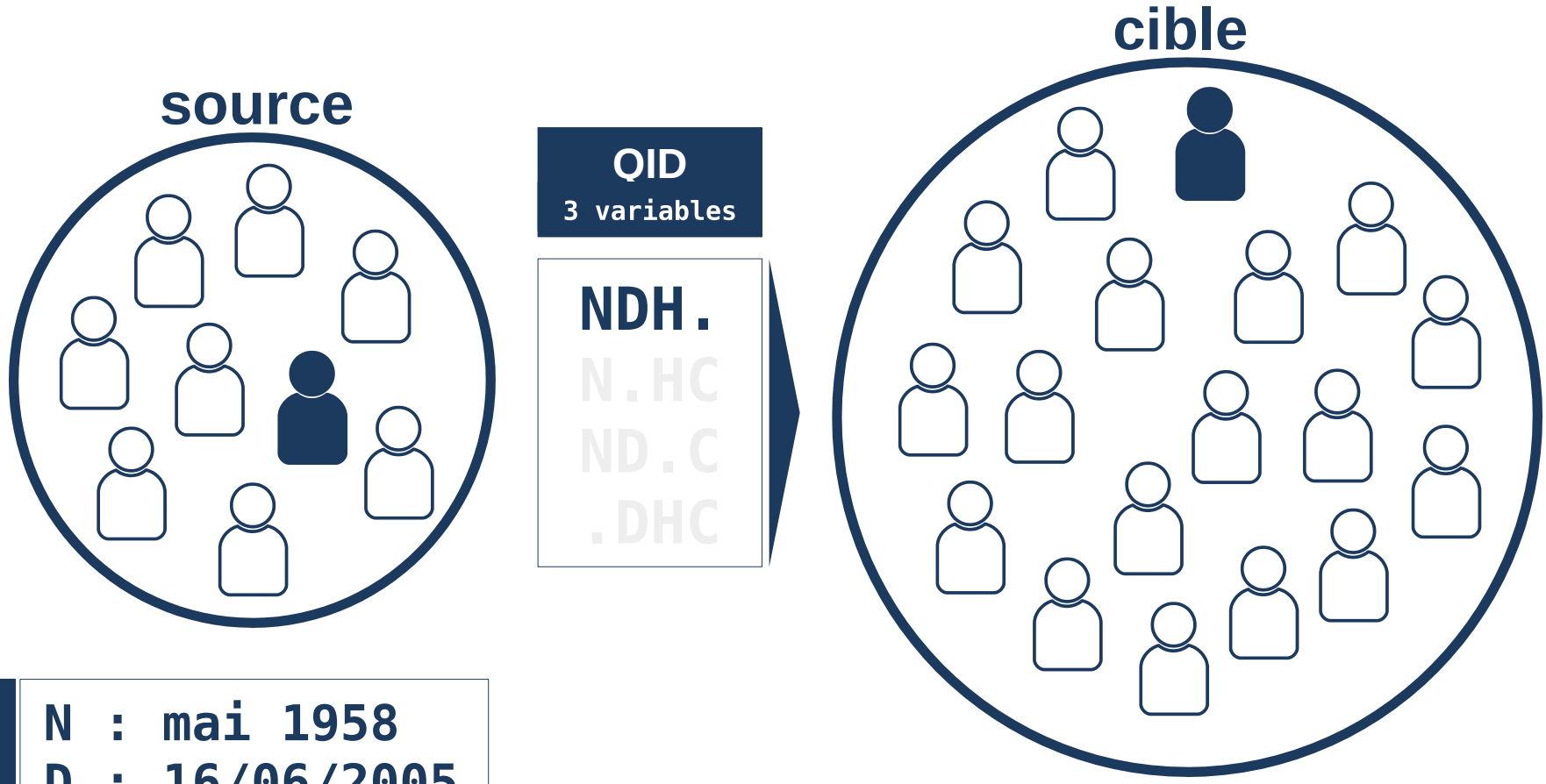
QID
N : mai 1958 D : 16/06/2005 H : 14/09/2003 C : 44100

Exemple d'appariement indirect

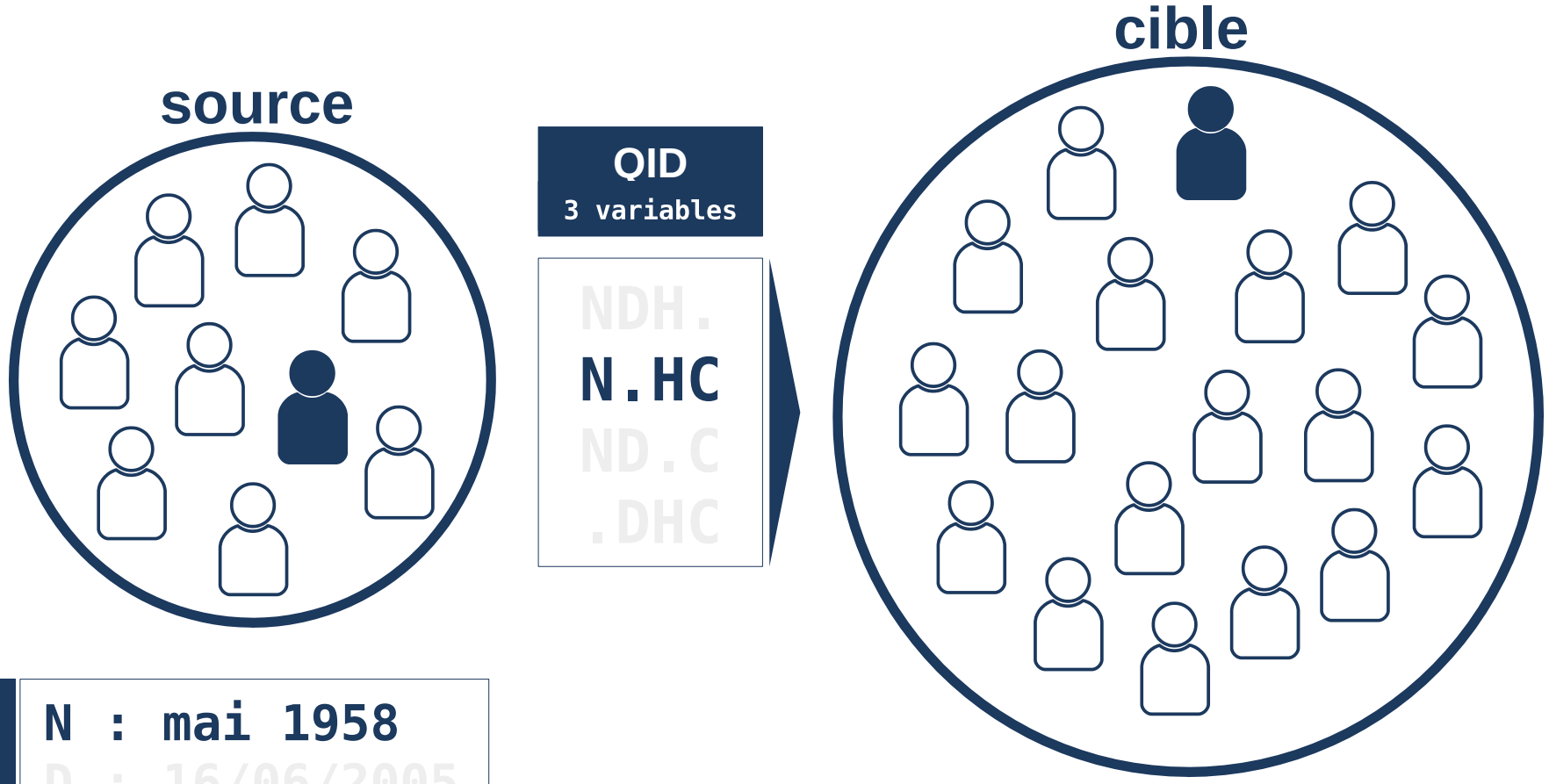


QID	
	N : mai 1958
	D : 16/06/2005
	H : 14/09/2003
	C : 44100

Exemple d'appariement indirect

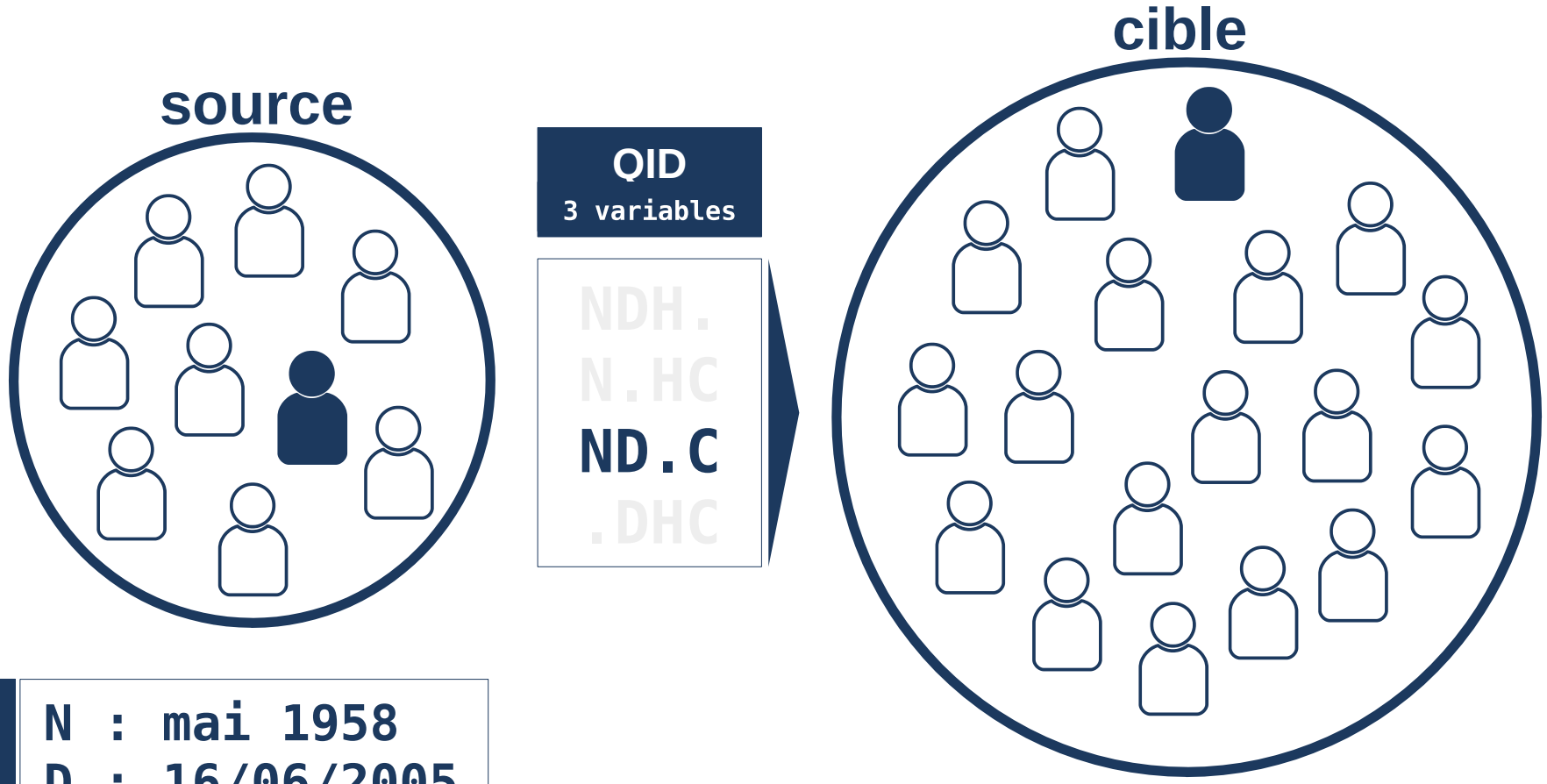


Exemple d'appariement indirect



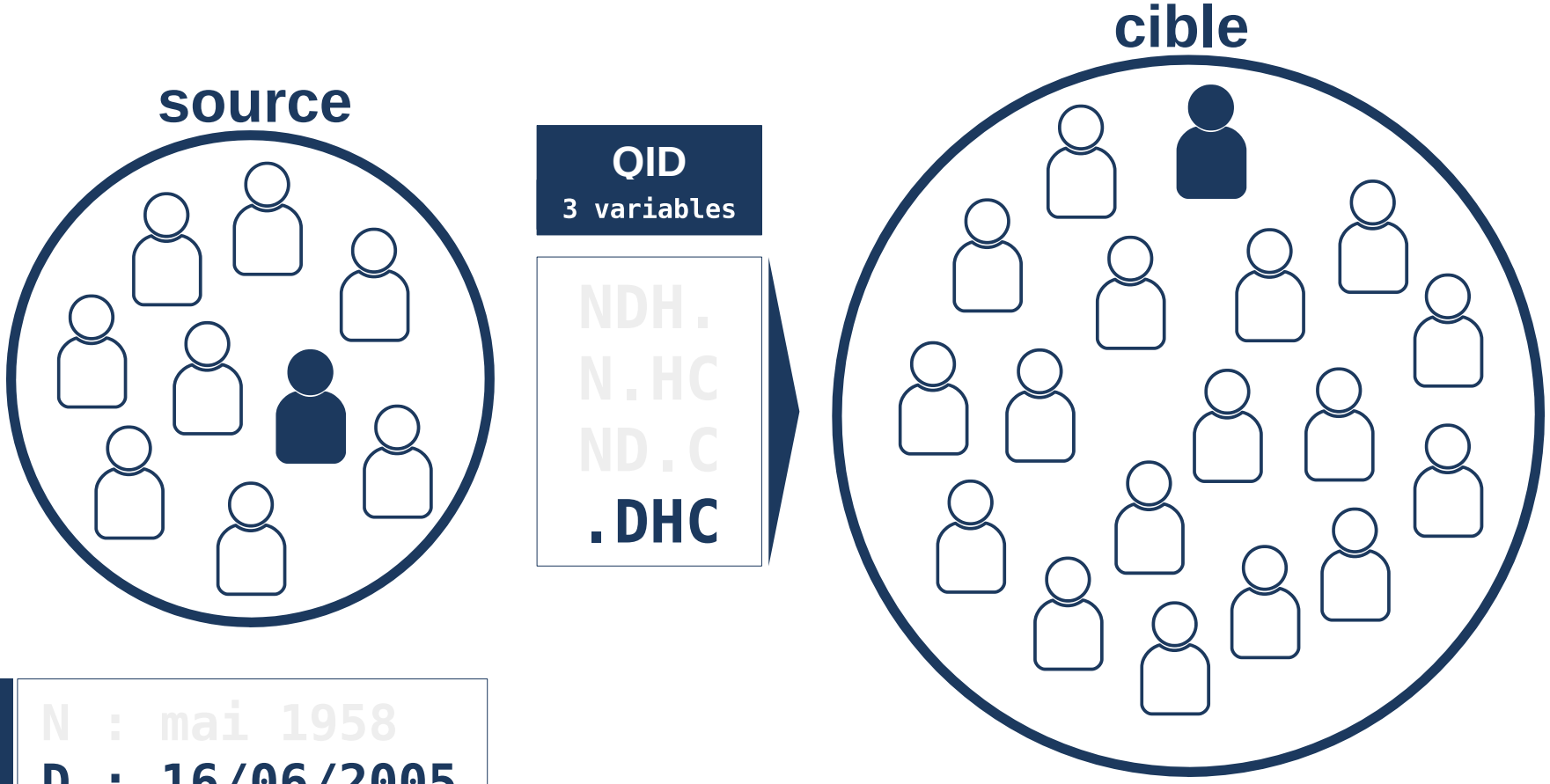
QID	
N :	mai 1958
D :	16/06/2005
H :	14/09/2003
C :	44100

Exemple d'appariement indirect



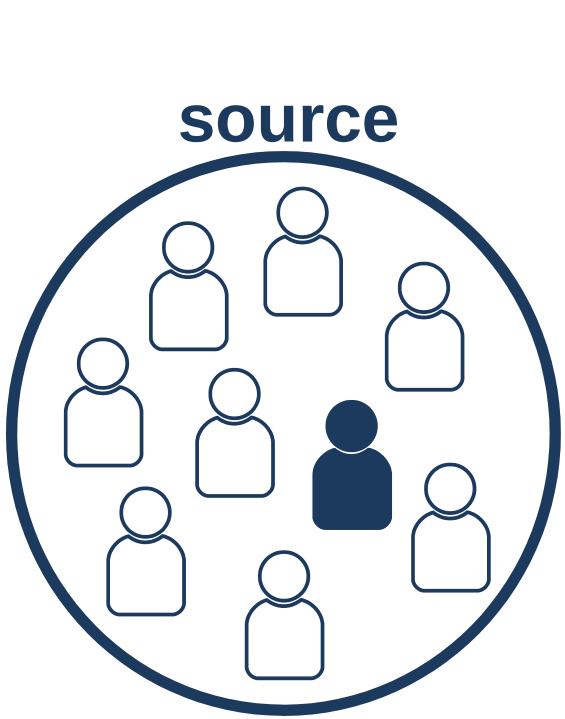
QID	
N :	mai 1958
D :	16/06/2005
H :	14/09/2003
C :	44100

Exemple d'appariement indirect



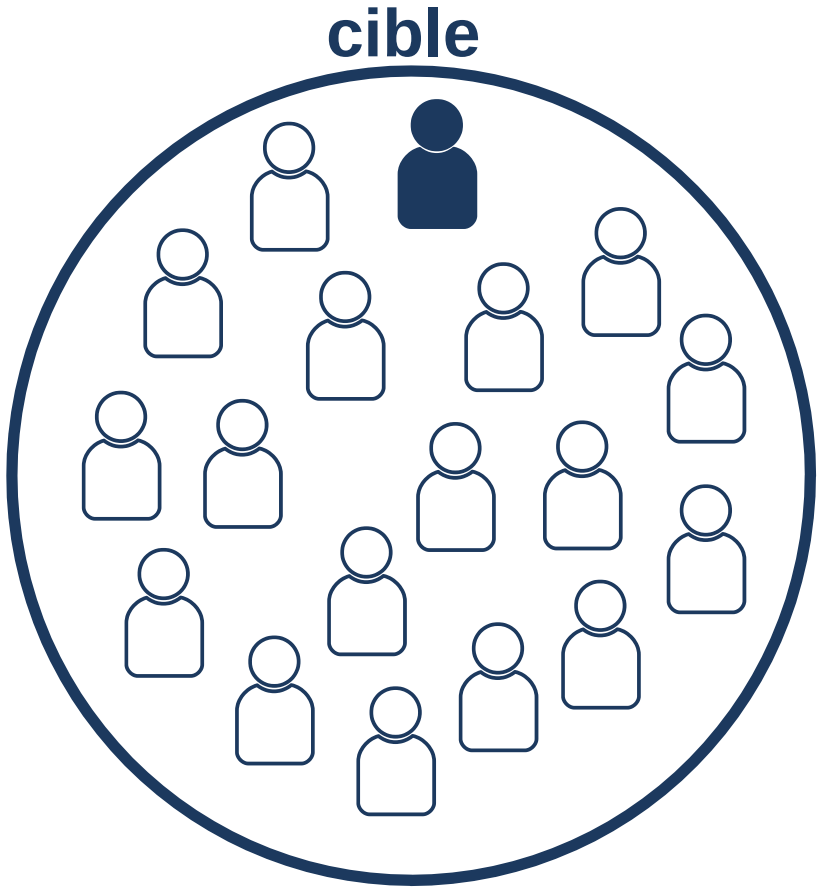
QID	
	N : mai 1958
	D : 16/06/2005
	H : 14/09/2003
	C : 44100

Exemple d'appariement indirect

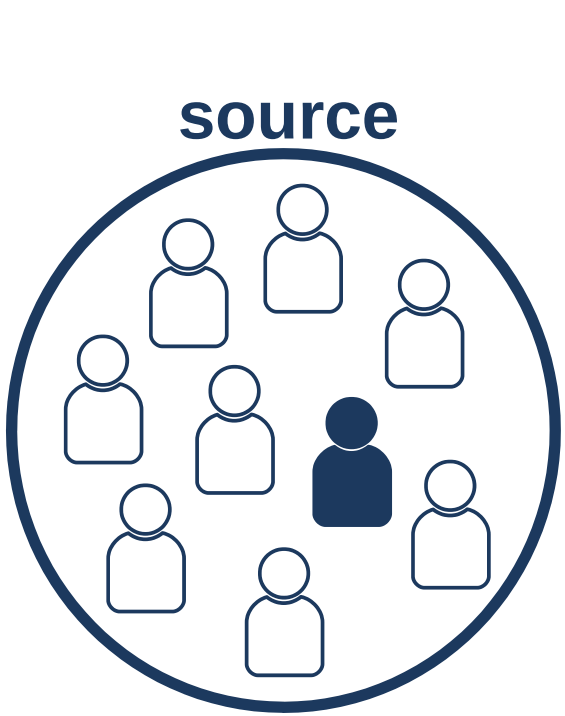


QID
N : mai 1958
D : 16/06/2005
H : 14/09/2003
C : 44100

QID 2 variables
N.H.
ND..
N..C
.DH.
..HC
.D.C

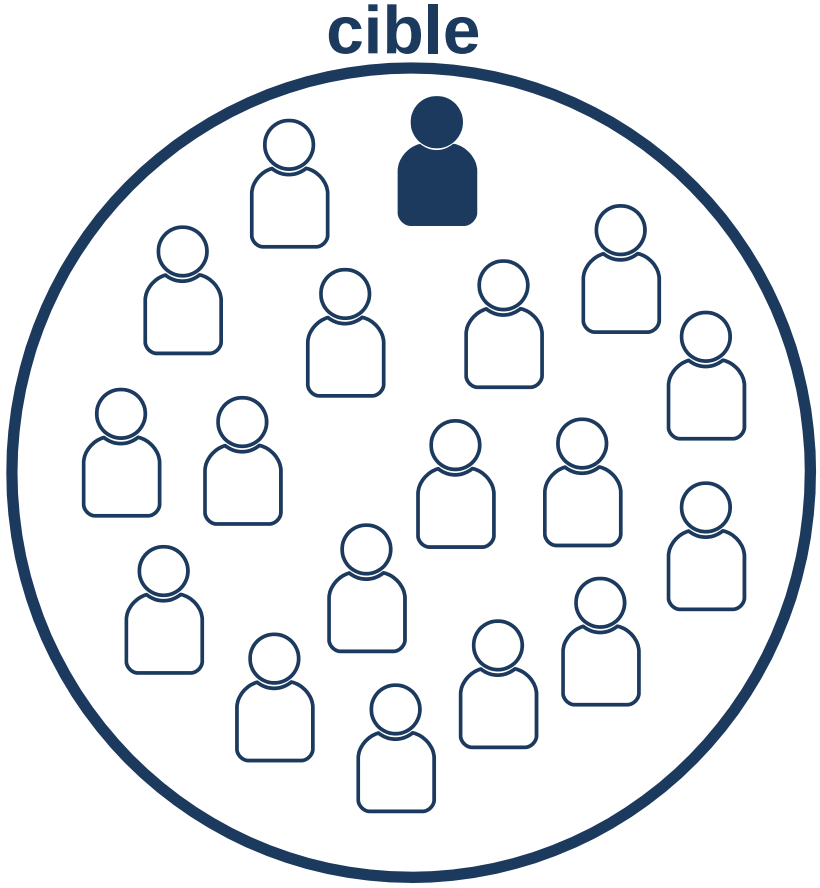


Exemple d'appariement indirect

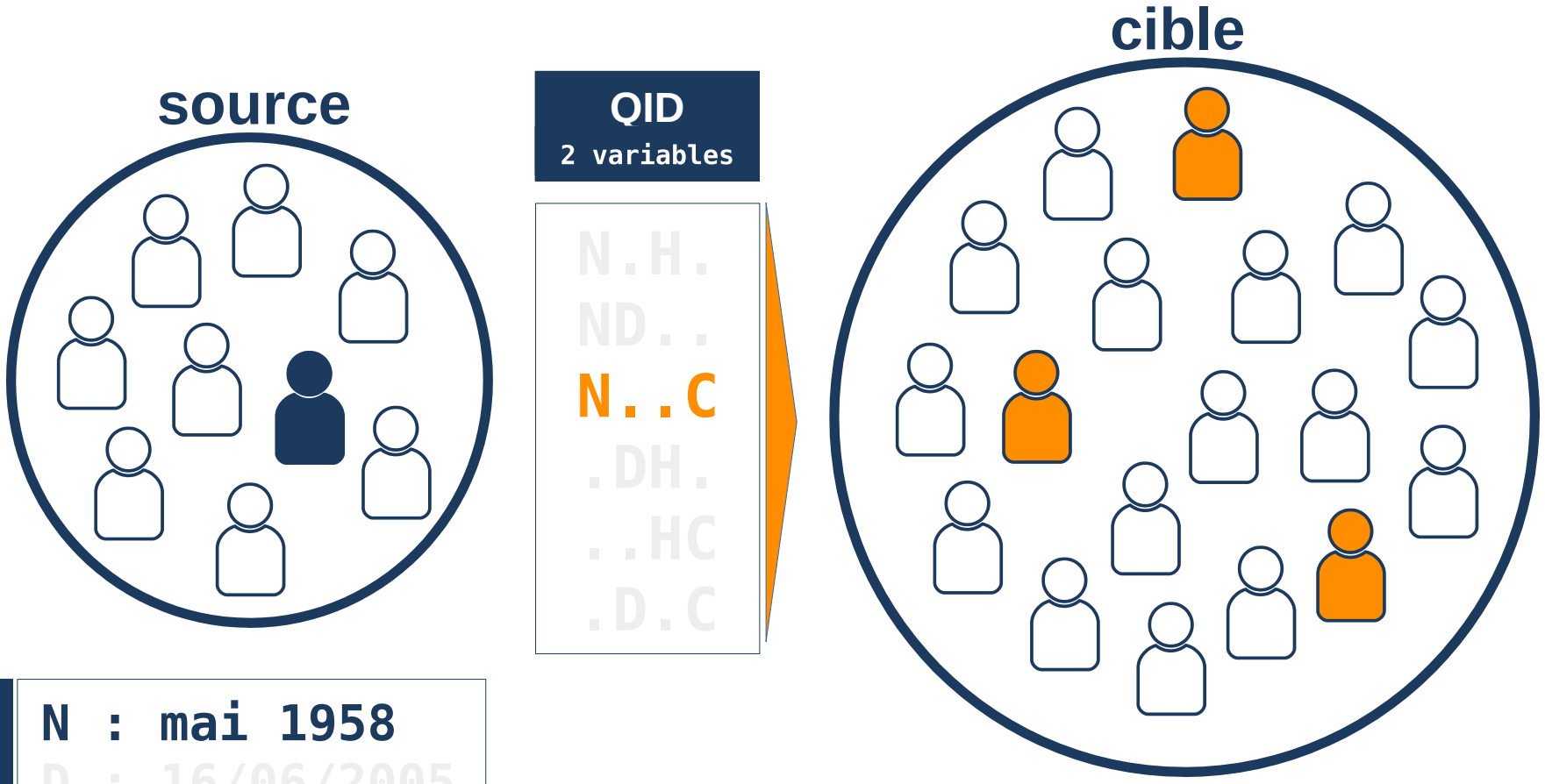


QID
N : mai 1958
D : 16/06/2005
H : 14/09/2003
C : 44100

QID 2 variables
N.H.
ND..
N..C
.DH.
..HC
.D.C

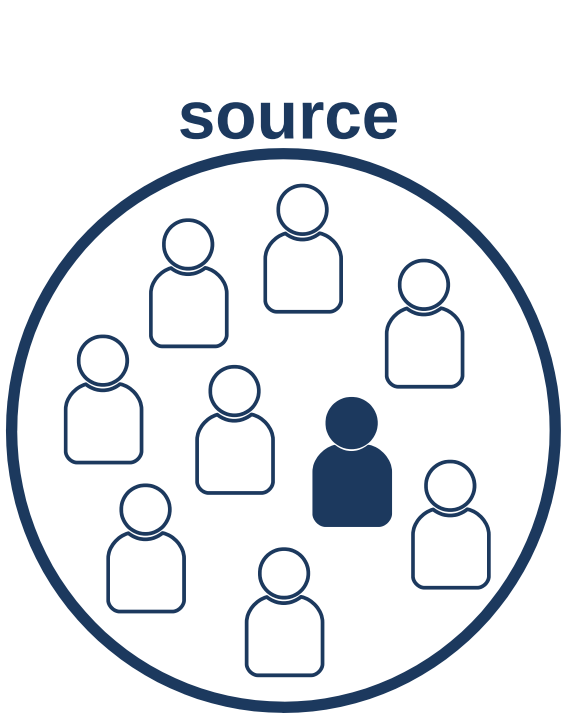


Exemple d'appariement indirect



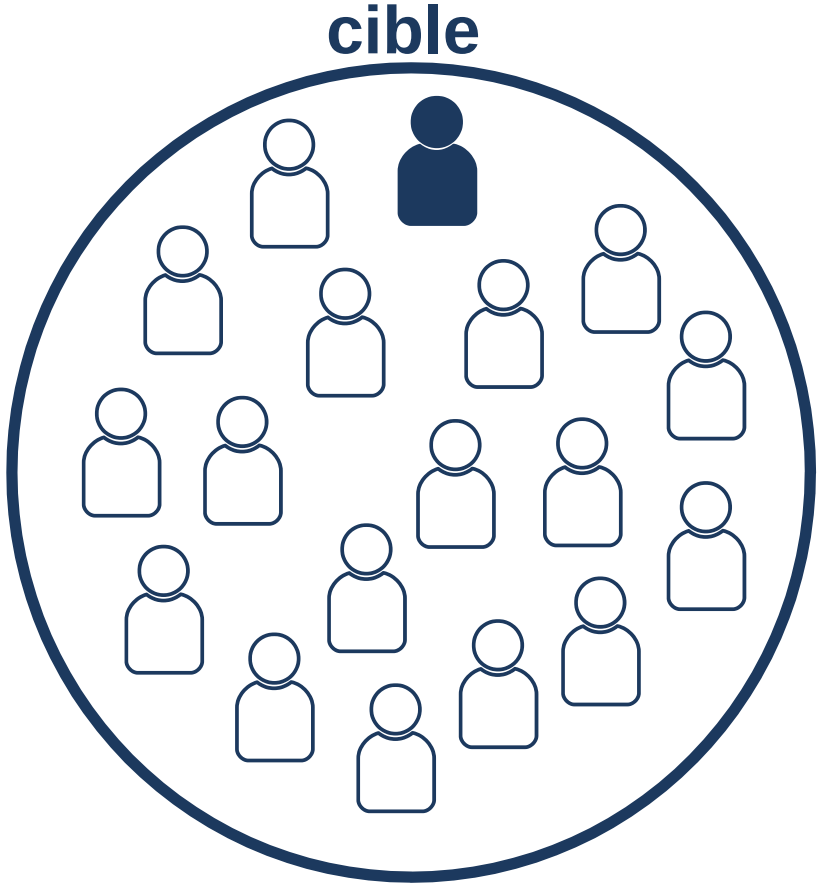
QID	
	N : mai 1958
	D : 16/06/2005
	H : 14/09/2003
	C : 44100

Exemple d'appariement indirect



QID
2 variables

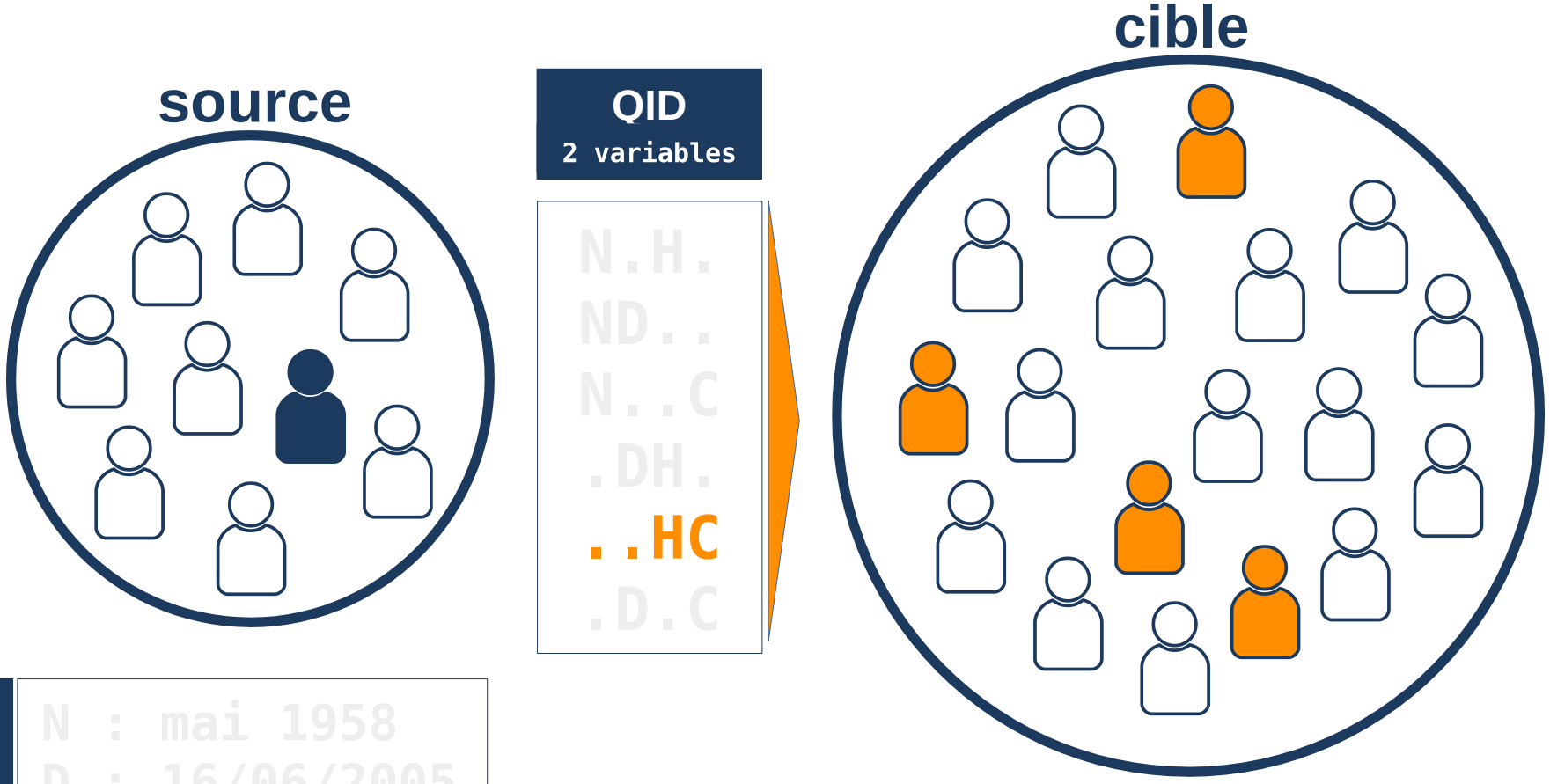
N.H.
ND..
N..C
.DH.
..HC
.D.C



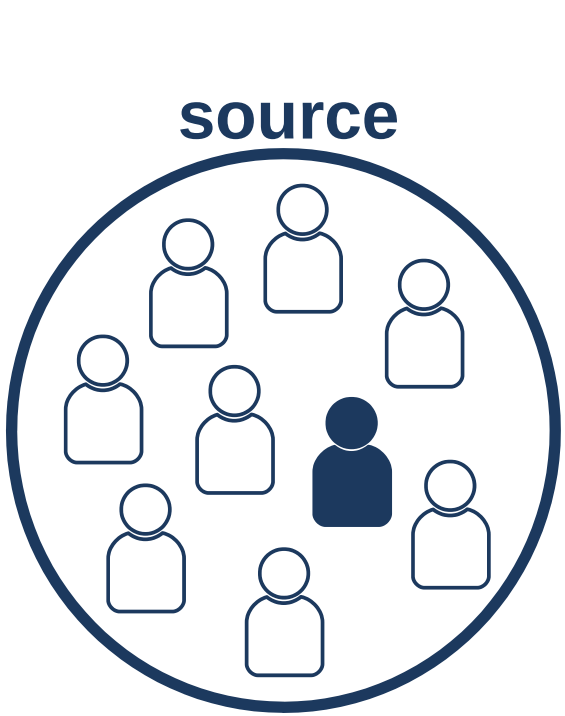
QID

N : mai 1958
D : 16/06/2005
H : 14/09/2003
C : 44100

Exemple d'appariement indirect

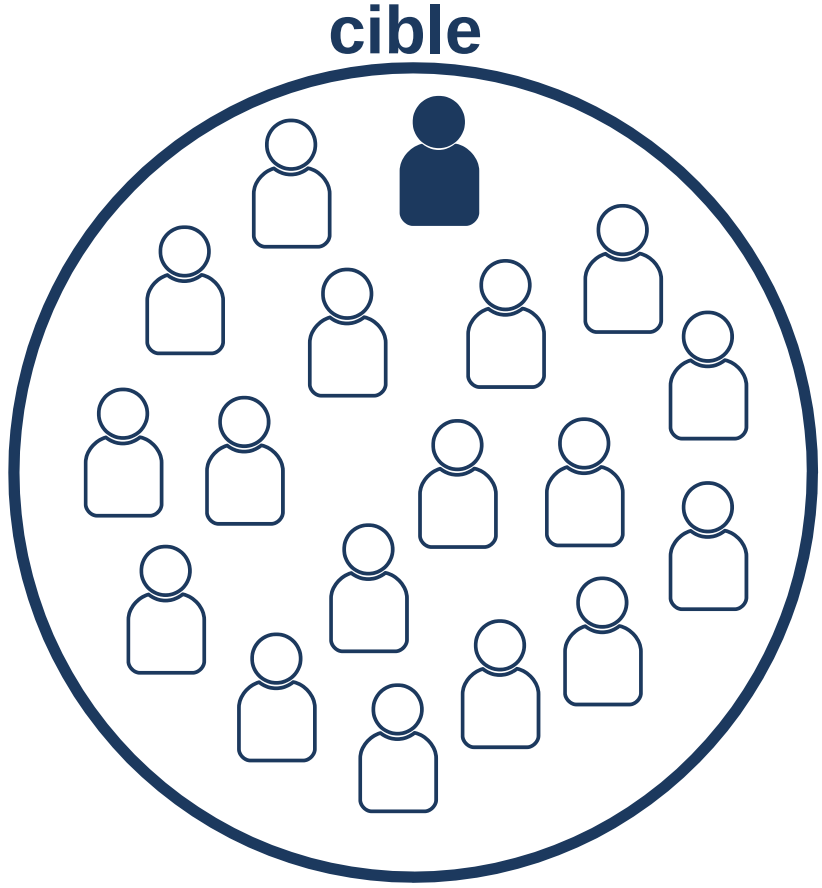


Exemple d'appariement indirect



QID
2 variables

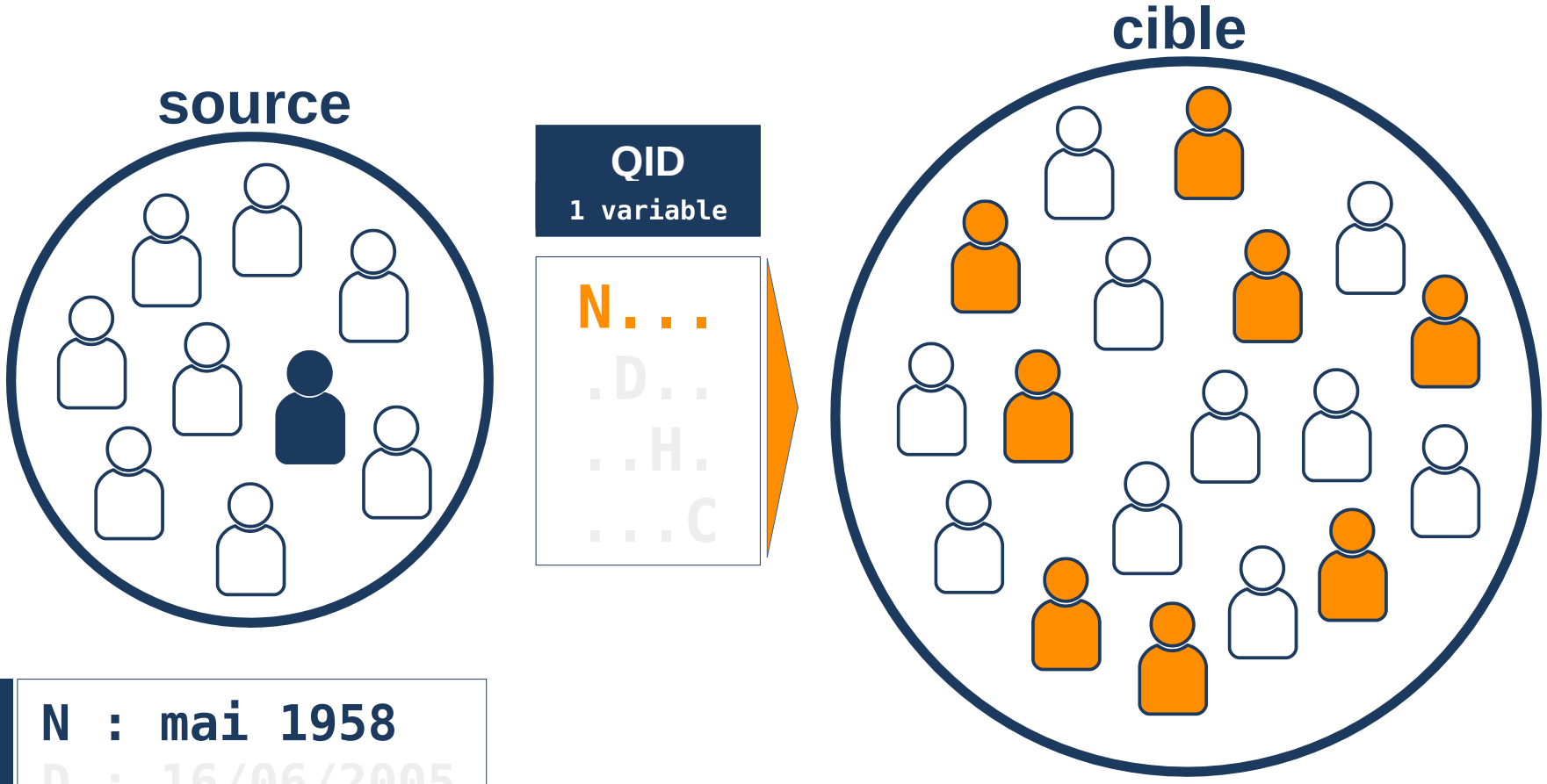
N.H.
ND..
N..C
.DH.
..HC
.D.C



QID

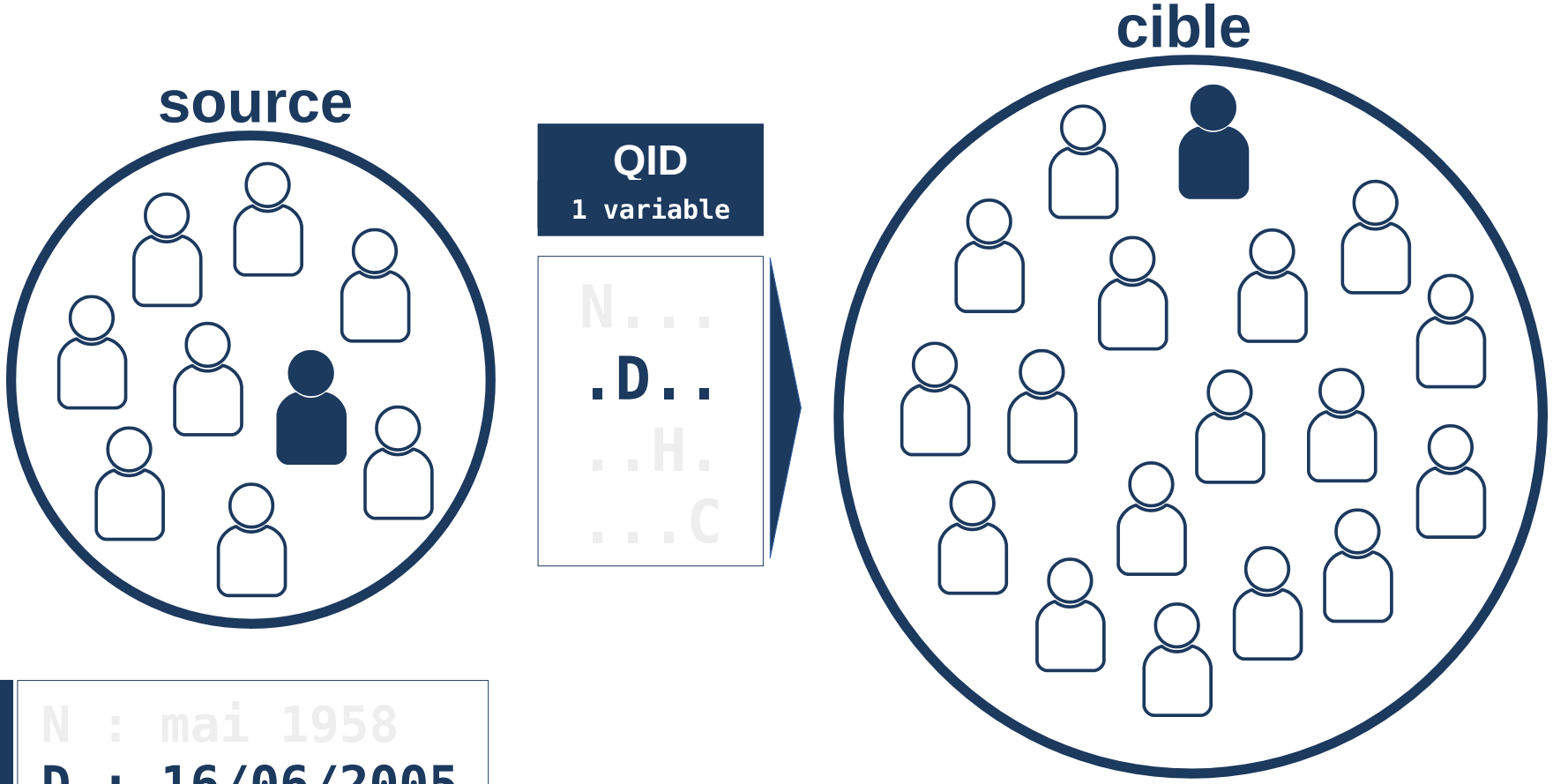
N : mai 1958
D : 16/06/2005
H : 14/09/2003
C : 44100

Exemple d'appariement indirect

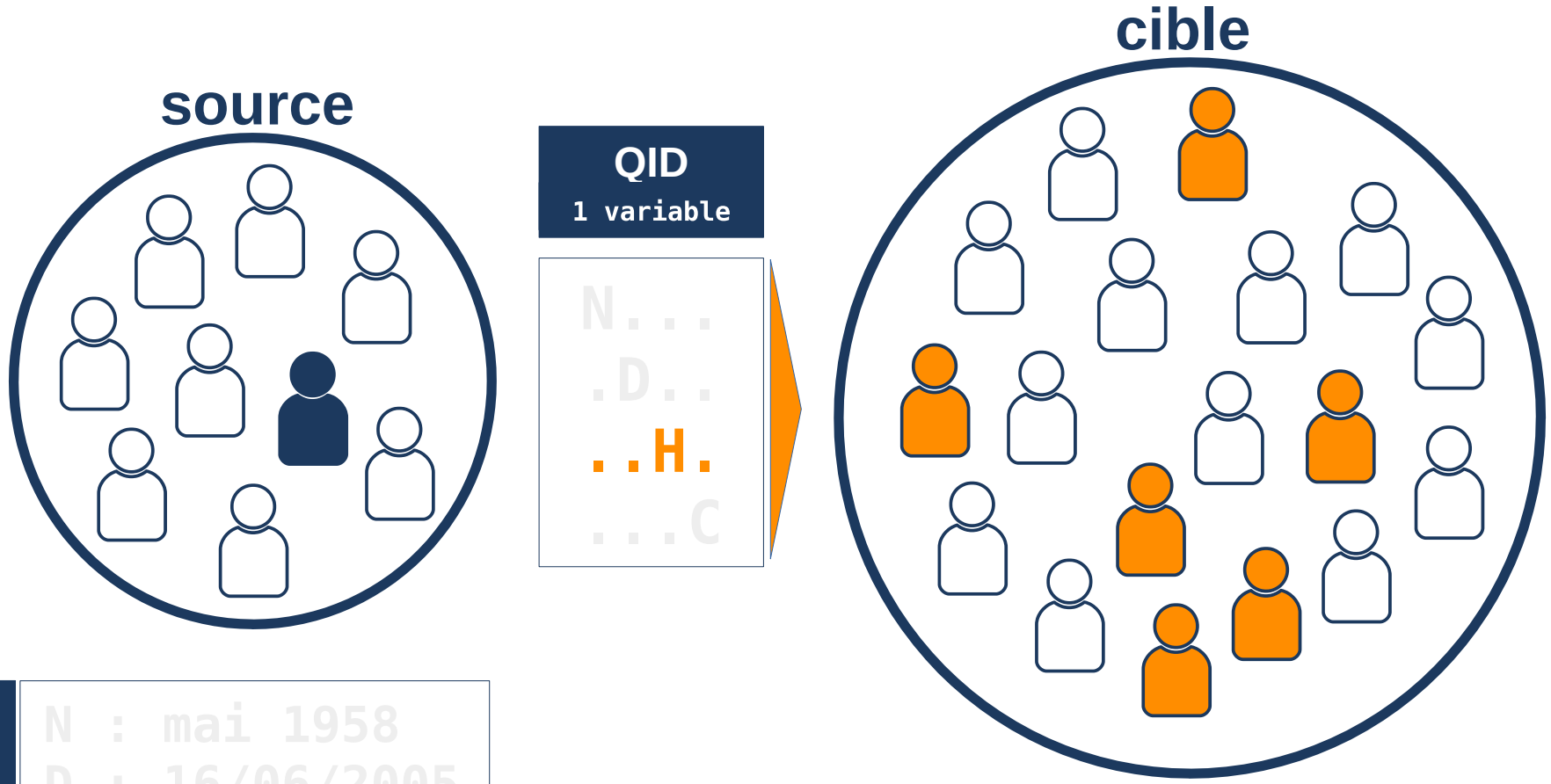


QID	
N : mai 1958	
D : 16/06/2005	
H : 14/09/2003	
C : 44100	

Exemple d’appariement indirect

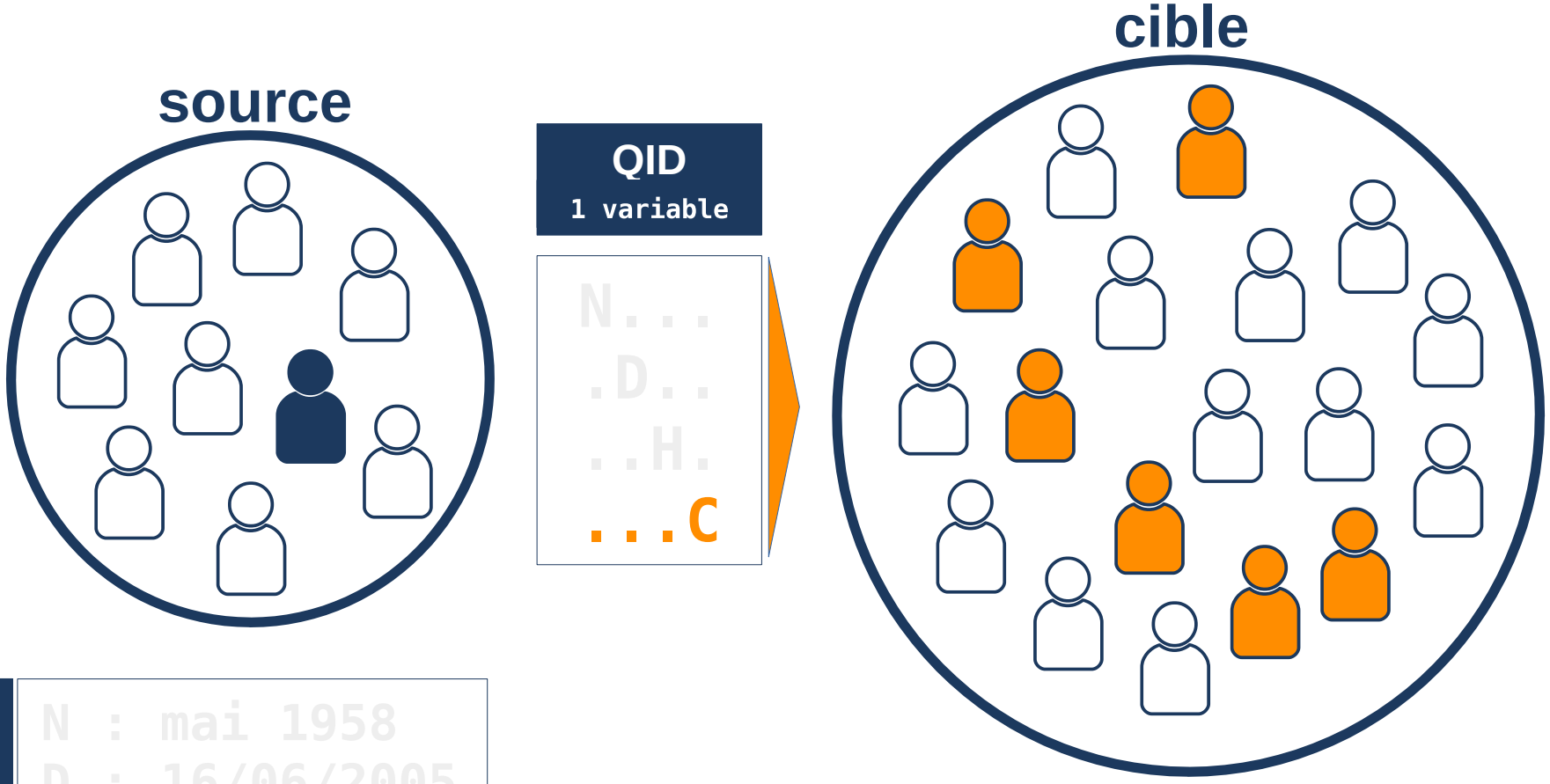


Exemple d'appariement indirect



QID	
	N : mai 1958
	D : 16/06/2005
	H : 14/09/2003
	C : 44100

Exemple d'appariement indirect

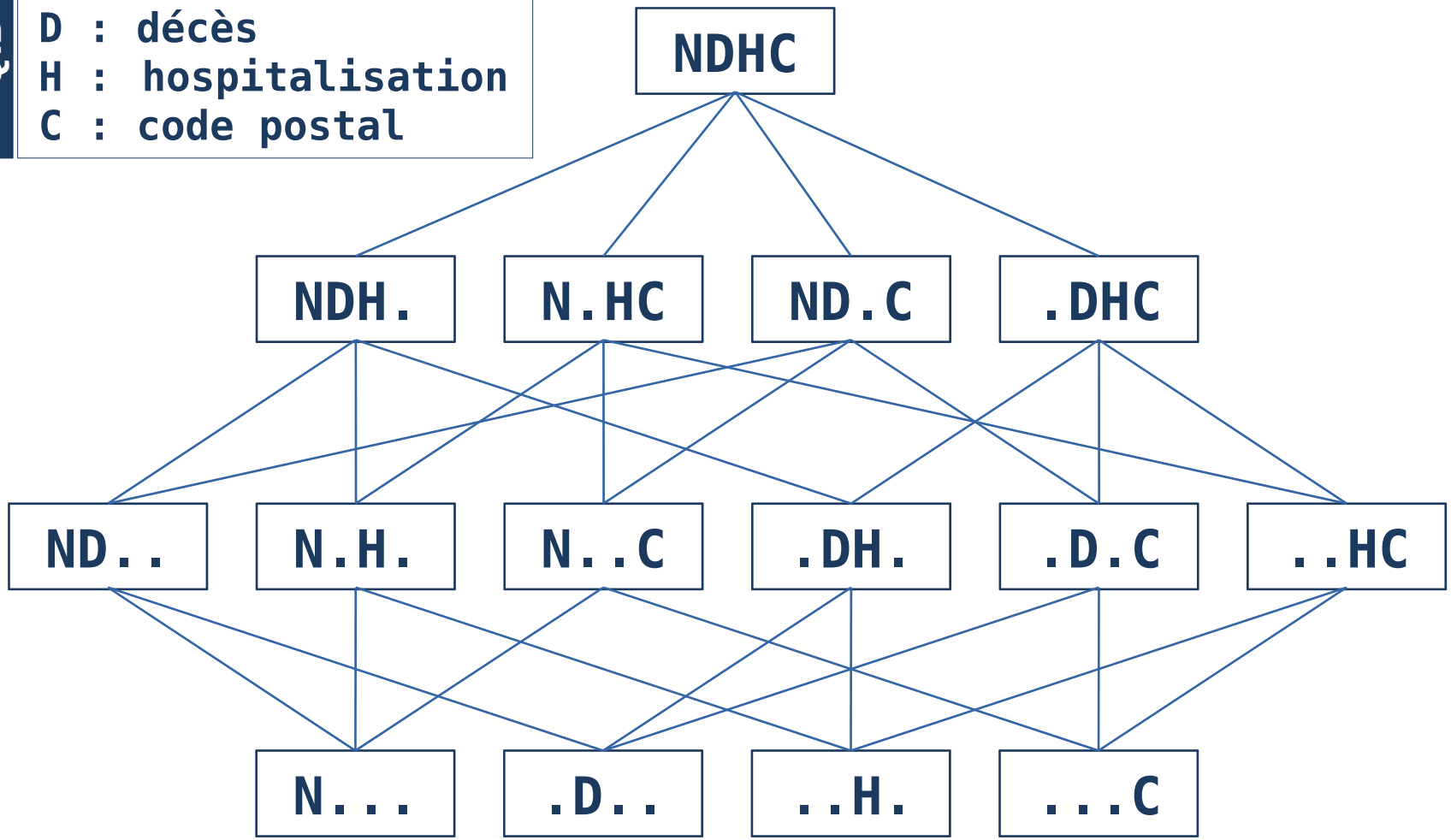


QID	
	N : mai 1958
	D : 16/06/2005
	H : 14/09/2003
	C : 44100

Exemple d'appariement indirect

QID

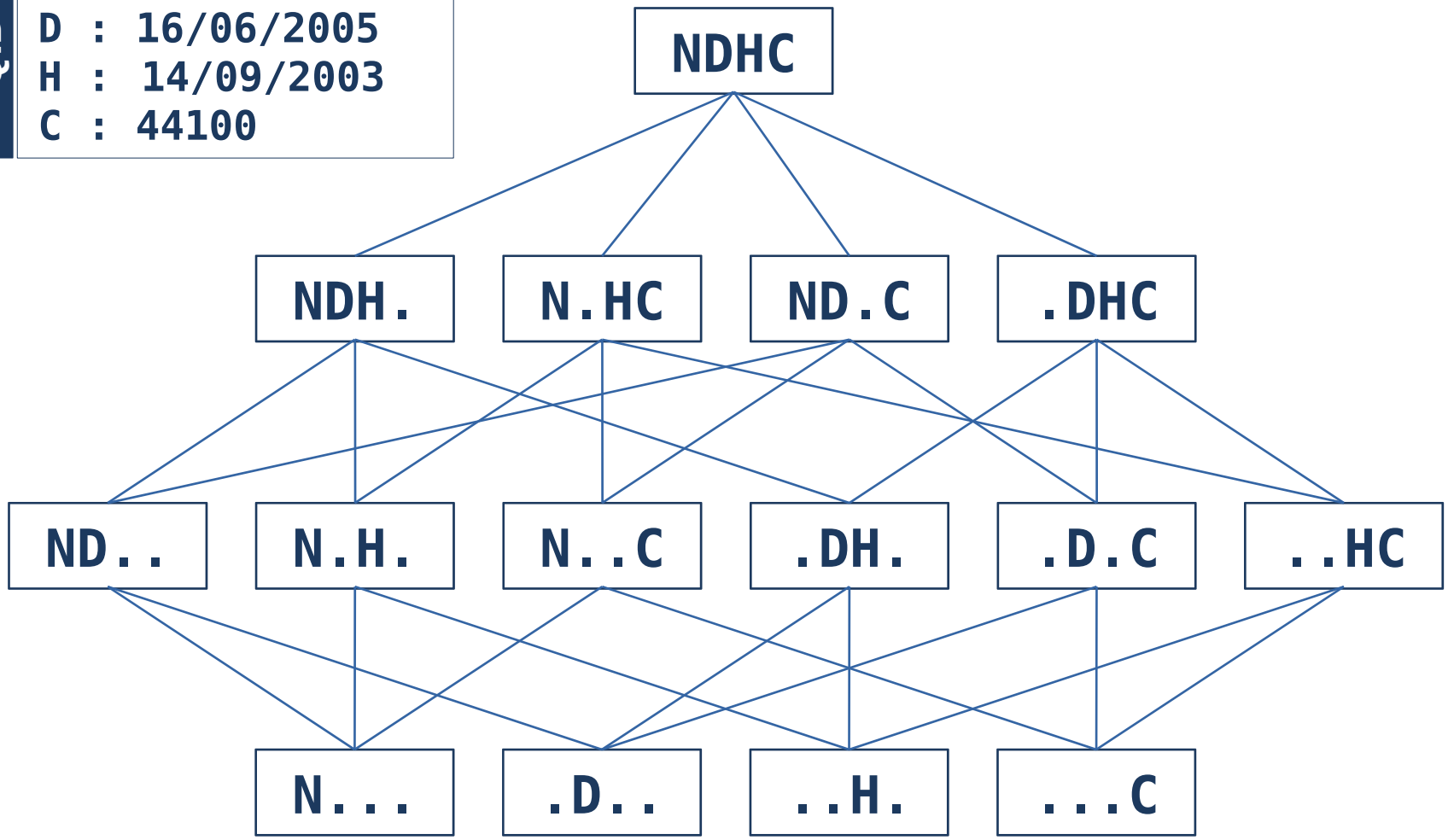
N : naissance
D : décès
H : hospitalisation
C : code postal



Exemple d'appariement indirect

QID

N : mai 1958
D : 16/06/2005
H : 14/09/2003
C : 44100

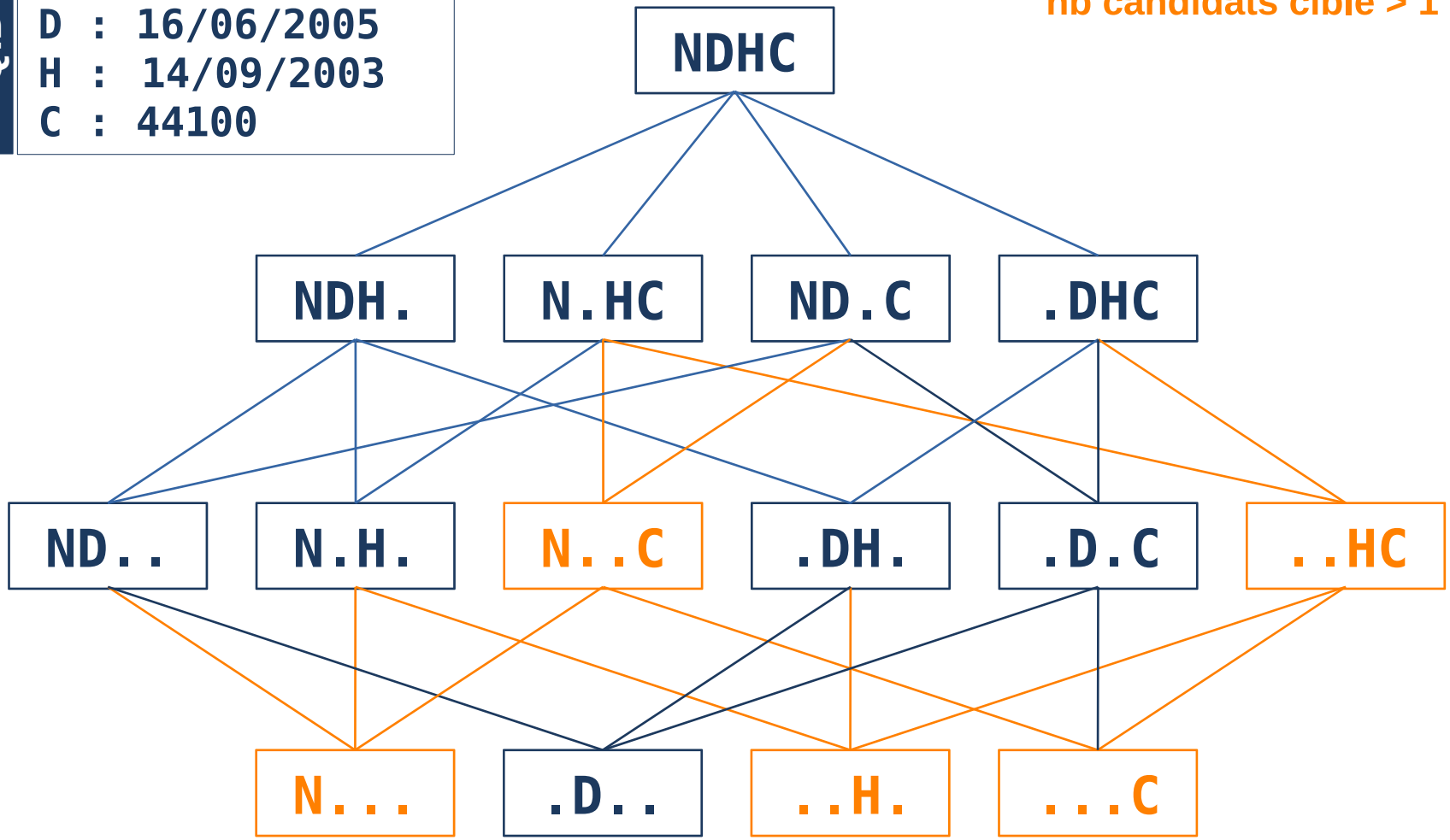


Exemple d'appariement indirect

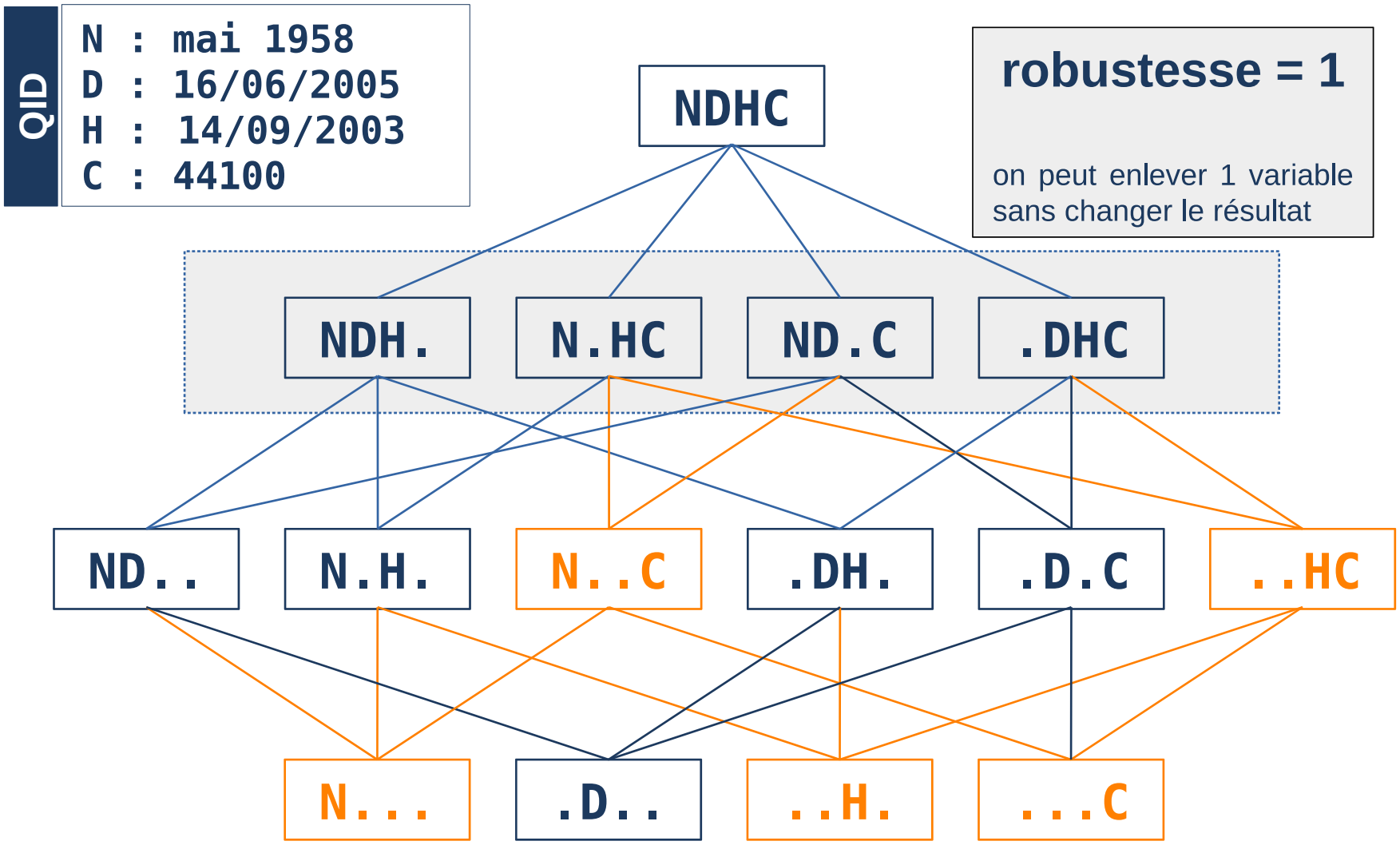
QID

N : mai 1958
D : 16/06/2005
H : 14/09/2003
C : 44100

nb candidats cible = 1
nb candidats cible > 1



Exemple d'appariement indirect



Temps de calcul

Exploration de toutes les combinaisons

- 2^N QIDs possibles

Complexité

- Complexité : $O(|S| . |C| . 2^N)$

Potentiellement long si réalisé naïvement

- $|S| = 30\,000$ $|C| = 3\,000\,000$ $N = 10$
- **Nombre d'opérations : 92 160 000 000 000**

Optimisations possibles

- Algorithmiques
- Structures de données efficaces

Caractéristiques de l'approche

Essayer tous les QIDs possibles

- Exploration totale de l'information
- Patient appariable → patient apparié

Résilience de l'approche face aux données

- Crucial pour des appariements avec le SNDS

Définition de la robustesse

- Métrique pour quantifier la confiance sur le résultat

Production d'une cartographie

- Métrique pour évaluer la pertinence du résultat

Quelques résultats…

Résultats obtenus

Issus de travaux communs entre
EA REPERES et Cubr

Mise en oeuvre de
la méthodologie & des algorithmes de Cubr

New metrics for assessing linkage quality in deterministic record linkage of health databases

Erwan Drézen *[1], André Happe[2], Sandrine Kerbrat[2], Frédéric Balusson[2], and Emmanuel Oger[2]

[1] *CUBR*

[2] *EA REPERES, University of Rennes*

Plateforme HAL

Appariement AVC Brest & SNDS

Project name	Brest Stroke Registry	
Patients number		
source	Registry	3985
target	SNDS	504795
ratio		0.007894
Linked pairs, n (%)		
≤ 0 missed var		3058 (76.7%)
≤ 1 missed var		3472 (87.1%)
≤ 2 missed var		3589 (90.1%)
all		**3803 (95.4%)**
Robustness		2.36
min-max		0 – 5
Linkage Variables		
number		9
missing data		4.5%
Runtime (seconds)		0.6

Une approche classique se situerait entre 76.7 % et 87.1 % seulement

Bonne robustesse globale

QID constitué de 9 variables

Appariement **Artome** & **SNDS**

Project name	Artome trial	
Patients number		
source	Trial	129
target	SNDS	22426
ratio		0.00575
Linked pairs, n (%)		
$\leq$ 0 missed var		61 (47.3%)
$\leq$ 1 missed var		102 (79.0%)
$\leq$ 2 missed var		120 (93.0%)
all		**127 (98.4%)**
Robustness		4.18
min-max		2 – 7
Linkage Variables		
number		12
missing data		7.8%
Runtime (seconds)		0.01

Une approche classique se situerait entre 47.3 % et 79.0 % seulement

Très bonne robustesse globale

QID constitué de 12 variables

Appariement **Artome** & **SNDS**

			robustness						
	ARTOME	SNDS	0	1	2	3	4	5	total
1	EUNnDddGS	.....		10	76	396	901	80	1463
2		....Ddd..	4	82	48				138
3		.U.....		1	20	50	10		81
4		.....S			13	49	7		69
5		....D....		1	18	23	2		44
6		.....G.		2	7	17	2		28
7		....Dd...		2	7	2			11
8		E.....GS	1	6					8
9		..Nn.....			7	1			8
10		....Ddd.S		2					8
11		EU.....G.		7					7
12		.U..D....			5	2			7
13		.U..Ddd..		2					4
14		E.....G.			2	1			3
15		....D...S		1	1	1			3
16		.U.....G.			1	1			2
17		.U.....S			2				2
18		..Nn....S		2					2
19		..N.....				2			2
20		.....GS			2				2
21		EU.....GS		1					1
22		.U..Dd...		1					1
23		..NnD....			1				1
24		....DddG.		1					1
25		....Dd..S		1					1
26	EUNn...GS	.....	3	707	661	2			1409
27		.U.....	16	53					74
28		.....S	26	35					61
29		.....G.		8					14
30		..N.....		2					2
31		E.....GS	1						1
32		E.....G.	1						1
33		.U.....S		1					1
34		.....GS	1						1

Extrait de la cartographie de l'appariement

Synthétise l'information de tous les QIDs possibles

Permet à l'expert d'exclure certains résultats si besoin

Projets d'appariement en cours

Registre « Sarcome » + SNDS

- **35 000** patients vs. **3 000 000 patients**
- Réalisé au sein du Health Data Hub

Registre « Sclérose en plaque » + SNDS

- **47 000** patients vs. **160 000** patients

Résultats bientôt disponibles

- Bons résultats
- Résilience de l'approche face aux données

Pour conclure…

Pour conclure

Appariement

Seul moyen de répondre à certaines questions

SNDS

« mine d'or » pour enrichir ses propres données

Mais NIR non disponible !

Appariement indirect via un QID

Définition du QID

En choisir un seul, quelques uns ***ou bien tous***

Pour conclure

Approche présentée	Approche classique
Industrialisation de la méthodologie	Développement « **ad hoc** »
Exploration **totale** de l'information	Exploration **partielle** de l'information
% d'appariement **maximisé**	% d'appariement **fluctuant**
Contrôle fin des résultats par l'expert	**Contrôle partiel** des résultats
Rapide même sur de gros volumes	Parfois **problèmes** sur gros volumes

Meilleurs %
Meilleure confiance

Merci de votre attention.

www.cubr.fr

contact@cubr.fr